Statistical Emulation for Environmental Sustainability Analysis

Thesis

Statistical emulation for environmental sustainability analysis

Oyebamiji Oluwole Kehinde, BSc., MSc., MPhil

A thesis submitted for the degree of
Doctor of Philosophy

Department of Environment, Earth & Ecosystems
The Open University
Walton Hall
Milton Keynes, UK
MK7 6AA

This project was funded by the ERMITAGE & The Open University

4^{th} December 2014

Date of Submission: 5 December 2014
Date of Award: 2 March 2015

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Abstract

The potential effects of climate change on the environment and society are many. In order to effectively quantify the uncertainty associated with these effects, highly complex simulation models are run with detailed representations of ecosystem processes. These models are computationally expensive and can involve computer runs of several days for their outputs. Computationally cheaper models can be obtained from large ensembles of simulations using a statistical emulation.

The purpose of this thesis is to construct cheaper computational models (emulators) from simulation outputs of Lund-Potsdam-Jena-managed Land (LPJmL) which is a dynamic global vegetation and crop model. This research work is part of a project called ERMITAGE. The project links together several key component models into a common framework to better understand how the management and interaction of land, water and the earth's climate system could be improved.

The thesis focuses specifically on emulation of major outputs from the LPJmL model; carbon fluxes (NPP, carbon loss due to heterotrophic respiration and fire carbon) and potential crop yields (cereal, rice, maize and oil crops). Future decadal changes in carbon fluxes and crop yields are modelled as linear functions of climate change and other relevant variables. The emulators are constructed using a combination of statistical techniques of stepwise least squares regression, principal component analysis, weighted least squares regression, censored regression and Gaussian process regression.

Further modelling involves sensitivity analyses to identify the relative contribution of each input variable to the total output variance. This used the Sobol global sensitivity method. The data cover the period 2001-2100 and comprise climate scenarios of several GCMs and RCPs. Under cross validation the percentage of variance explained ranges from 52-96% for carbon fluxes, 60-88% for the rainfed crops and 62-93% for the irrigated crops, averaged over climate scenarios.

Acknowledgments

I am most grateful to my supervisors Dr. Neil Edwards, Prof. Paul Garthwaite and Dr. Philip Holden for their guidance. They are the best supervision team. They provided me with invaluable research and written directions that made this thesis exciting to write. I sincerely thank you for all your effort and assistance throughout my PhD. They are highly acknowledged.

I am also indebted to the entire staff of Department of Environment, Earth & Ecosystems. The services provided by the office staff members Helen Philip and Liz Lomas have been great, you have done so much for me. I want to say thank you. There are, of course, many other people who have made this research work possible that I would like to thank. Barbara Pizzileo, the ERMITAGE project manager, I appreciate your kind support and contribution. I would like to thank Dr. Sibyll Schaphoff and Dr. Dieter Gerten for providing the simulation data used in this thesis. I appreciate the entire people that participated in the ERMITAGE project especially, Fred, Marc, Santosh, you are warmly appreciated. My colleagues in the office have been fantastic and full of fun, and we have had a great wonderful time together during our PhD. I have to say thank you for Frazer, Kate, Peter, Alice, Harley, Adele and Chris.

I cherish the unquantifiable support and assistance rendered by my wife Oluwakemi Oyebamiji and my daughter Blessing Oyebamiji. They understood the pressure of the work and continually encouraged me even during very difficult time. Your deep love, tender care, cute attention and unparalleled desire to make me happy during my PhD are greatly appreciated. Thank you for all the love.

I thank the wonderful friends who have helped me along the way, particularly Akinyele Kazeem, Orelesi Emmanuel, Dr. Segun Ogundimu and Aluko Olalekan. I appreciate the forum created by our friendship and the benefit I have derived from it. I love you all. My appreciation will remain incomplete if I do not mention the moral and spiritual supports, advice and prayer of Pst. Kehinde Olajide, Pst. Mike Bright, Pst. Emmanuel Amadi, brother Lucky, Mrs. Obi, Mrs. Wunmi Adediran, Pst. Festus, Pst. Joel and entire members of MFM Milton Keynes branch.

Parts of this thesis have been published, presented and abstracted

Publications

Oyebamiji, O.K., Edwards, N.R., Garthwaite, P.H., Holden, P.B., Schaphoff, S. and Gerten, D. (2015). Emulating Global Climate Change Impacts on Crop Yields. *Statistical Modelling*, 1471082X14568248, first published on January 18, 2015, doi:10.1177/1471082X14568248,http://smj.sagepub.com/content/earl

Oyebamiji, O.K., Edwards, N.R., Garthwaite, P.H., Holden, P.B., Schaphoff, S. and Gerten, D. Gaussian Process Emulation of Impact of Climate Change on Crop Yields. In preparation for submission to *Environmental & Ecological Statistics.*

Oyebamiji, O.K., Edwards, N.R., Garthwaite, P.H., Holden. LPJmL Emulator. Submitted as a part of Emulation and Coupling Report for EU deliverable 3.2. http://ermitage.cs.man.ac.uk/?q=node/52.

Conference Talks

Oyebamiji, O.K., Edwards, N.R., Garthwaite, P.H., Holden. Statistical Emulation for Environmental Sustainability Analysis. *ERMITAGE 2nd Annual Meeting and Workshop, Potsdam, Germany*, Sept. 2012. *http : //ermitage.cs.man.ac.uk/sites/default/files/public − merged.pdf*

Conference Posters

Oyebamiji, O.K., Edwards, N.R., Garthwaite, P.H., Holden. A New Emulator of Global Climate Change Impacts on Crop Yields. In *Uncertainty in Computer Models 2014 Conference, Sheffield*, July 2014. *www.mucm.ac.uk/UCM2014/Forms/Abstracts/ContributedPosterOluwole.pdf*

Oyebamiji, O.K., Edwards, N.R., Garthwaite, P.H., Holden. Predicting Terrestrial Biospheric Response to Climate Change. In *Uncertainty in Complex Models (UCM) 2012 Conference, University of Sheffield.* http://mucm.ac.uk/UCM2012/Forms/ContributedPosterOyebamiji.pdf

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The global climate is changing and this has a great negative impact on ecosystems. It adversely affects the environment, life quality, the economy, agriculture. Any change in the climate system over time, whether due to natural variability or human activity, is referred to as climate change (IPCC, 2007). Of a particular priority in this thesis is the assessment of the impacts of future climate change on terrestrial ecosystems. Assessment of climatic impacts on global vegetation is attracting increasing attention among modellers, since modelling provides a thorough way of quantifying the biospheric response under climate uncertainty. Adequate knowledge of the impacts of climate change on crop yield is required to fully understand global food security.

Future increases in the demand for food and reduced availability of land for growing pose major challenges for human society as a result of rising projections for world population over the next century. Growth in demand for food will also put pressure on water resources. Rising temperature, CO_2 emissions and associated climate change will affect global food and water supplies. Potential climatic impacts on vulnerable populations through lack of adequate food and regular water supply have been identified as serious threats.

Modelling tools for studying the interaction between climate change and its effects are relatively under-developed. In spite of tremendous progress in the statistical analysis of climate change studies, quantifying the climatic impact

on vegetation, agriculture and water resources still suffer problems. Moreover, adequate knowledge of climatic impacts on terrestrial ecosystems is important. For instance, comprehensive and detailed understanding of the impact of climate change on crop yields is required in order to maintain good management practices for sustainable agriculture. However, this assessment is challenging and computationally intensive models have to be run.

Empirical quantification of the reduction of climate change impacts in different sectors by moving from a no mitigation approach to several mitigation scenarios was the major focus of Warren et al. (2013) and Arnell et al. (2013). Based on the quantitative evidence from their studies, they opined that urgent global measures can prevent the larger impacts of climate change that are projected to occur by mid-century.

Several modelling approaches have been used for the global assessment of climatic impacts on vegetation, and in particular relating agricultural crops to weather. Some authors used popular ecosystems process-based models eg Bondeau et al. (2007); Fader et al. (2010) to simulate detailed physical and biological processes. Specifically, Leemans & Solomon (1993) implemented a water balance model within a GIS that integrates databases, while Fischer et al. (2001) extended the approach with the Global Agro-Ecological Zones (GAEZ) model.

Similarly, Rosenzweig et al. (2014, 1994); Parry et al. (2005, 1999) and Iglesias et al. (2000) evaluated global consequences of various climate change scenarios on crop productivity with a process-based approach. Recently, Muller & Robertson (2014) compared the performance of two global crop models, LPJmL of Bondeau et al. (2007) and Decision Support System for Agrotechnology Transfer (DSSAT) of Jones et al. (2003) for projecting future crop productivity and its integration for economic assessment (Nelson et al., 2014).

In contrast to the process-based methods, other authors have relied solely on empirical or statistical methods which utilize correlative relationships between variables without a detailed description of physical behaviours. Simulated and

historical data have been explored to assess the global climatic impact on crop yields in Lobell & Field (2007); Sacks et al. (2010); Schlenker & Roberts (2009); Lobell & Burke (2010). Empirical models are often linked to process-based models, while process-based models also incorporate some empirical information. Most global climate impact assessments have combined these two approaches. Another major concern is the issue of projected uncertainty. There are many sources of uncertainty in projected climate changes, for instance, deficiencies in modelling key processes that regulate the important biophysical effects like water and carbon cycles. Boundary conditions for different global climate models and regional climate variability can introduce uncertainty in impact assessment (Christensen et al., 2007). Nevertheless, generating ensembles of simulations has provided useful means of quantifying the uncertainty in projections of regional climate changes (Graham et al., 2007; Beniston et al., 2007).

Uncertainty associated with a complex computer model is often quantified by running a computer model (simulator) as many times as possible. However, because computer models require large amounts of computer time to run, this is not always possible; see Sacks et al. (1989) for a comprehensive discussion of computer experiments. In order to overcome this difficulty, a reliable and efficient technique has to be developed based on a statistical representation of the simulator. One solution to this problem is to construct an emulator from limited simulator runs, which will help in lowering the computational burden. Emulation has been applied in various areas of climate science (Warren et al., 2008; Holden et al., 2010a,b; Arnell et al., 2013; Gosling et al., 2007; Heyder et al., 2011; Lobell & Burke, 2008).

The goal of this thesis is to assess the global impact of climate change on natural vegetation and agricultural crops. The Lund-Potsdam-Jena-managed Land (LPJmL) model is a process-based dynamic global vegetation and crop model. The global responses of carbon and vegetation patterns under climate change for both natural and agricultural ecosystems can be simulated by LPJmL. LPJmL is

also capable of simulating water runoff because of the close connection between global carbon and the water cycle. The model is used in this study to simulate the response of terrestrial ecosystems to climate change. In other words, it predicts the behaviours of complex dynamics associated with interactions between vegetation, climate and landuse. The run times for LPJmL are high and this often makes the full assessment of the impact of climate change on the ecosystems impossible. In addition, uncertainty and sensitivity analyses are difficult to perform and can be impractical because of the number of runs that are required (further details of LPJmL and its simulations is described in section 4.3).

Statistical emulation can help to overcome this problem. Emulation integrates both the process-based model and statistical techniques. Besides, emulators offer rapid and relatively quick alternatives for projection of climate change impacts on agricultural productivity for diverse climate scenarios. For instance, Warren et al. (2013); Arnell et al. (2013) considered both economic and physically based modelling approaches to project global and regional impact of climate change, but due to computing constraints only a limited set of scenarios were investigated. Another benefit of emulation is the provision of a measure of uncertainties associated with the projections.

This thesis describes the development of novel statistical modelling techniques for the emulation of major output responses from the computationally expensive simulator LPJmL. Our emulation approach is relatively straightforward and easy to apply. We demonstrate the techniques in the emulation of carbon fluxes (net primary productivity, fire carbon, carbon loss due to heterotrophic respiration). We then emulate potential crop yields (cereal, rice, maize and oil crops). We also perform sensitivity analysis of these emulation results.

We used the following statistical techniques; ordinary least squares (OLS) regression, weighted least squares (WLS), censored regression (CR), and principal component analysis (PCA) and Bayesian regressions. Some methods, for instance, Bayesian regressions, are not directly applied to the original data from

LPJmL, but instead they are applied to the residual outcome from OLS results. In particular, Gaussian Process (GP) emulation is limited by the ensemble size because of the inversion of the covariance matrix (Biegler et al., 2011).

In this thesis, we have considered both one-stage and two-stage strategies. In stage 1, we applied both OLS and censored regression to the multi-dimensional data. In stage 2, emulator residual analyses are performed where we explore spatial interpolation of the unexplained residuals from stage 1. The two-stage approach provides substantial efficiency gains and improves the predictive capability of the emulator over the standard methods.

Emulation approaches

In this thesis, the following four methods for emulating the LPJmL outputs will be explored

1 A one-stage method using only least squares regression (OLS).

2 A one-stage method using only censored regression (CR).

3 A two-stage method using combinations of OLS, PCA and WLS. Here, OLS is used for the analysis in the first stage while WLS and PCA are used in the second stage.

4 A two-stage method using a combination of OLS, PCA and GP regression. OLS is again applied to the first stage modelling while GP regression implements the second stage. PCA was used to reduce the dimensionality of the residuals from the first stage and the resulting components (PC scores) are the inputs for the second stage.

Ermitage project

The work reported in this thesis is part of the project ERMITAGE (Enhancing Robustness and Model Integration for The Assessment of Global Environmental

Change). An aim of the ERMITAGE project was to produce methods of coupling models of the natural and socio-economic systems. Linking key component models through a common framework enables a better understanding of how the management and interaction of land, water and the Earth's climate system could be improved.

One of the objectives of the project is to couple the climate and impact models together by feeding in outputs from a climate system model (MAGICC6) to a global dynamic vegetation and crop model (LPJmL). However, computational speed is highly problematic for coupling complex models together. The good news is that with a statistical emulation the computational limitation can be significantly reduced. Emulation would also facilitate other calculations (eg sensitivity analysis) that would not be practical using the LPJmL model directly. Here, we are interested in investigating the interaction between climate, natural systems (terrestrial biosphere impact) and agricultural land-use. The work reported in this thesis is an important part of the project. The main objective of the work in this thesis is to model the global terrestrial biospheric response to climate change and anthropogenic emission by constructing statistical emulators.

Thesis outline

The structure of this thesis is as follows. Chapter 2 presents background on the ERMITAGE project and its objectives. It describes the concept of emulation. It provides a brief discussion of the natural carbon fluxes and notable agricultural crops selected for emulation in this thesis. Chapter 3 gives comprehensive reviews of the relevant literature on climate change and terrestrial biosphere, carbon fluxes and crop yields. It also discusses common emulation techniques before focusing on censored and Bayesian regression including multivariate GP.

Chapter 4 gives a basic description of the impact models and simulation data used for the analyses in this thesis. The models described are MAGICC6, ClimGen and LPJmL. Chapter 5 focuses on the methodology and, in particu-

lar, the statistical methods of OLS, WLS, PCA, GP and CR are fully described. Chapter 5 also discusses variable selection criteria for stepwise regression, measurement of model performance and sensitivity analysis.

Chapters 6, 7 and 8 present the detailed procedures for emulating carbon fluxes and crop yields with their respective results. Chapter 9 contains discussion, concluding remarks and suggestions for future work.

Chapter 2

General background

This chapter gives some general background needed for this thesis. The purpose of emulation is outlined in section 2.1 and section 2.2 describes the general background of the climate system, climate change and its effects on terrestrial ecosystems. Quantities related to carbon fluxes are defined. Section 2.3 covers the crops whose yields are modelled in this thesis and outlines their importance. Section 2.4 outlines the future climate scenarios that are considered in this thesis.

2.1 Statistical emulation

An emulator or a metamodel is a statistical representation of a complex function that is able to stand in for a model in applications where the model or where running many instances of the model would be too expensive to evaluate. Emulation is a tool for simplification of models that leads to reduced-form representations of complex models that is computationally much faster, potentially smoother functionally, and hence easier to couple to other models. It is a statistical framework for predicting the output from a complex deterministic function, such as a computer model (O'Hagan, 2006). Emulators can be used for prediction, uncertainty and sensitivity analyses as well as parameter calibration.

Generally, an emulated function f of inputs u can be related to a regression

analysis, where

$$f(u) = \sum_j \beta_j h_j(u) + \mathbf{Z}(u) \tag{2.1}$$

and β_j comprises uncertain coefficients, h_j is specified regressor functions, and $\mathbf{Z}(u)$ is a residual process.

Computer models (simulators) are used as approximations for complex physical experiments. Simulation runs are often expensive, and in some instances can take days or weeks. The simulator is not a complete representation of reality because of the presence of model discrepancy. A simulator takes inputs $\mathbf{U} = u_1, \ldots, u_n$ and produces outputs $\mathbf{y} = f(\mathbf{U})$. The discrepancy between the output $\mathbf{y}$ and true real world value, $\mathbf{z}$, gives rise to an error and it could be from either the input variables or model structure (structural error).

Uncertainty analysis involves the identification of the important uncertainty in model structures or parameters and quantitative estimation of their uncertainty or variability in model components. Uncertainty in a computer model comes in various forms: parameter uncertainty, model inadequacy, residual variability, parametric variability and measurements error. For instance, lack of perfect representation of key processes involved in a model may result in a structural deficiency, eg error inherent in the Earth system model processes, while imperfect information on the true parameter values in a model could give rise to parametric error. For instance, parameter values can be obtained from the existing literature and expert opinions or estimated from models which could be subjected to imprecision as a result of varying from one context to another.

Uncertainty can also be categorized as intrinsic and epistemic uncertainties. Intrinsic uncertainties are common in those phenomena that are subjected to variability or randomness in nature (eg climate variability). It is often difficult to reduce this type of uncertainty. Epistemic uncertainty, on the other hand, occurs as a result of insufficient knowledge of the system being studied or through scarcity of data. Epistemic uncertainty arises through the lack of a full understanding of the causes and effects of system processes. This type of uncertainty can be

reduced by gathering more data and increasing the knowledge base (Kennedy et al., 2001; Monod et al., 2006).

The Bayesian technique (details in Chapter 5) incorporates uncertainty modelling in its parameter estimations. Posterior mean estimates are produced by the simulator for any unsampled observations (predictions), together with posterior variances that measure error (uncertainty) in the predictions. More details of emulation as a surrogate for complex computer models can be found in Sacks et al. (1989) and O'Hagan (2006).

2.2 Carbon fluxes

The carbon cycle distributes carbon between the land, ocean and atmosphere. Atmospheric CO_2 concentration was stable at 260-280 *ppm* before 1750 (preindustrial). After that period, the CO_2 level rose significantly to around 380 *ppm* in 2005. This increase in CO_2 concentration is attributed to human activities like the burning of fossil fuels and deforestation.

Anthropogenic CO_2 emissions are a major factor causing climate change. CO_2 emission plays significant roles in the natural carbon cycle through continuous flows of large amounts of carbon among the ocean, the terrestrial biosphere and the atmosphere. The process of photosynthesis allows atmospheric carbon to be converted to plant biomass. Terrestrial plants absorb CO_2 from the atmosphere; plant, soil and animal respiration release carbon back to the atmosphere in the form of CO_2. Vegetation fire is another significant source of CO_2 emission to the atmosphere (IPCC, 2007; Cockell et al., 2007; Ciais et al., 2013; Le Quere et al., 2013).

Adequate knowledge of the global carbon cycle is important in understanding feedbacks between the biosphere and atmosphere and is thus essential in the analysis of global climate change. Quantifying uncertainty associated with carbon fluxes is needed to guide policy and management decisions. Besides, the global carbon cycle and its interaction is a dynamic and complex phenomenon. Under-

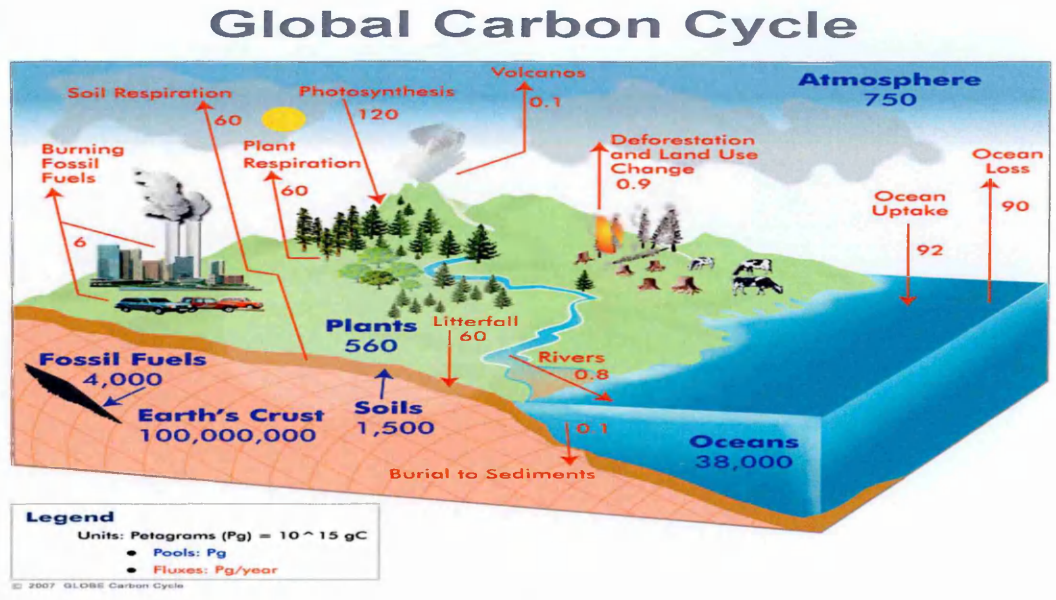
standing it will help us predict how atmospheric climate will change in the future. Major carbon fluxes result from net primary productivity, heterotrophic respiration and carbon loss due to fire activity. These three fluxes drive the terrestrial carbon cycle, so if we want to really understand and make reliable projections of atmospheric CO_2 concentrations, we need to understand and quantify fluxes into and out of our natural vegetation including crops, forests and pasture.

The Figure 2.1 below is a representation of the carbon cycle showing different biogeochemical processes by which carbon is exchanged among the biosphere and atmosphere of the Earth.

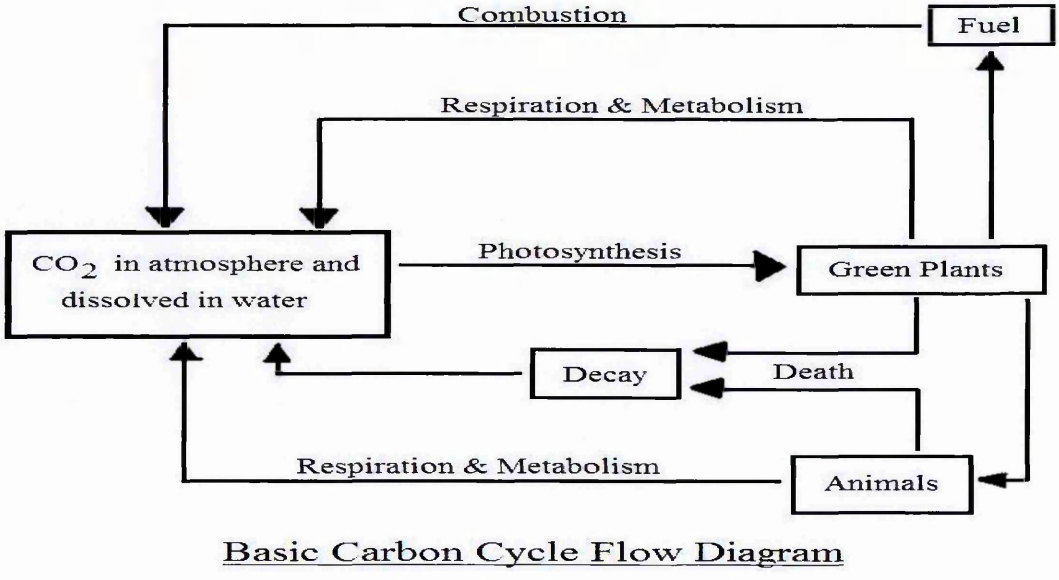
2.2.1 Definition of Net Primary Productivity (NPP)

Net Ecosystems Production (NEP) is the difference between the rate of production of living organic matter (NPP) and the decomposition rate of dead organic matter, so it is the net accumulation of carbon by ecosystems (Watson et al., 2000). NPP is the rate at which vegetation in an ecosystem fixes carbon from the atmosphere minus the rate at which it is returned by the plants themselves (McGuire et al., 1993). NPP is a major component of NEP. NPP is the net carbon gain by vegetation through photosynthesis and plant respiration (for maintenance and growth), and it represents the primary source of food for Earth's heterotrophic organisms. NPP is the dominant process in the biospheric carbon budget (Box, 1988). There is a considerable change in pattern and trend of NPP in response to climate change and anthropogenic CO_2 emission (McGuire et al., 1993; Melilo et al., 1993; Peng et al., 1995). NPP causes seasonal variations in atmospheric CO_2 (Keeling et al., 1996) and is a measure of plant productivity and crop yield (Milner et al., 1996). Net Biome Production (NBP) is the change in carbon stocks after carbon losses due to natural or anthropogenic disturbances like fire are accounted for.

The NPP, NEP and NBP are useful tools for quantifying the impact of land transformation in global change research and represent the initial input of carbon



(a)



(b)

Figure 2.1: Global carbon cycle. Source from www.wrsc.org/attach-image/global-carbon-cycle

to the biosphere (Imhoff et al., 2004). NPP, NEP and NBP are dominant processes in the biospheric carbon budget (Box, 1988). The equations below relate the three carbon concepts together.

$$NPP = GPP - R_a \tag{2.2}$$

$$NEP = NPP - HR \tag{2.3}$$

$$NBP = NEP - FC \tag{2.4}$$

where R_a is the autotrophic respiration (plant respiration), while Gross Primary Production (GPP) is the total amount of carbon fixed by plants during photosynthesis and HR is the heterotrophic respiration and FC is the fire carbon (Kirschbaum et al., 2001).

2.2.2 Definition of Heterotrophic Respiration (HR)

HR is the microbial decomposition of organic carbon, so it is the carbon lost by organisms in ecosystems other than through plants. It constitutes the respiration by animals that live above ground and by all organisms that live in the soil and litter layer, including fungi and other decomposer organisms. Heteroptrophic respiration also includes the carbon released in the decomposition of standing dead trees and coarse woody debris. It can also be defined as the release of CO_2 during the process of decomposition of organic matter in the soil by soil animals and other decomposer organisms (Kirschbaum et al., 2001; Raich & Schlesinger, 1992).

2.2.3 Definition of Fire Carbon (FC)

Disturbance from fire triggers ecosystems change and is one of the significant factors for composition, structure and dynamics of vegetation. Disturbance is the non-uniform and irregular occurrence of destruction of vegetation structure either by anthropogenic or natural means. It is a sudden change in environmental conditions that causes a pronounced change in an ecosystem with great consequence. It plays an integral part in shaping the global vegetation (Thonicke et al., 2001; Whitlock et al., 2003).

Fire tends to perform a similar role to heterotrophic respiration in the global carbon cycle. It reverses the photosynthesis process by transforming and decomposing carbohydrates back to carbon emissions, water, and energy, with the

amount of biomass consumption and carbon release depending on the degree of severity. Fire significantly affects the biogeochemical and carbon cycles. About 3.9 Giga tonne (Gt) of carbon are released annually into the atmosphere through biomass burning (Andreae, 1991; Goward et al., 2008). Thonicke (2005) examined changes in vegetation productivity, global fire regimes and related trace gas emissions. Wildland fires have influenced the global carbon cycle and its complex interaction with climate determining vegetation distribution. Changes in global climate have the potential to increase the incidence and severity of fire (Sommers et al., 2014; Bowman et al., 2009; Dillon et al., 2011).

Fire is a major agent in global vegetation disturbance regimes and the ecological processes that determine the distribution of biomes (Kelley et al., 2014; Yue et al., 2014). Furthermore, it can change landscapes from carbon sinks to carbon sources (Loehman et al., 2013) while interaction between fire and climate could cause a reduction in ecosystems diversity (Mitchell et al., 2014). It is important to study the influence of fire on the dynamic equilibrium and potential changes in vegetation at the global scale. Sound knowledge of the relationships, interactions and feedback effects between climate and fire regimes is required to account for the importance of fire in the carbon cycle under climate change.

2.3 Crop yields

The following agricultural crops: temperate cereal, rice, maize and oil which is a combination of soybeans, rapeseed, sunflower and groundnut from LPJmL model are emulated in this study. There will be a slight change in definition of oil crop under section (7.2) of Chapter 7. We believe it might not be appropriate to group both rapeseed and groundnut together because functionally rapeseed is quite different from groundnut and therefore might respond quite differently to climate change. For instance, rapeseed requires vernalization (period of cold winter temperature) for its flowering. They are also grown under quite different baseline climate conditions. Therefore, we shall model groundnut as a separate

crop under section (7.2) of Chapter 7.

2.3.1 Temperate cereal

Temperate cereal is a monocotyledon grass cultivated for the edible components of its grain. Temperate cereal grains are a globally significant staple food. They have high nutritive value and are a rich source of vitamins, minerals and carbohydrate. Major examples are wheat, rye, oats and barley.

Wheat

Wheat is a cool season crop widely spread around the world. It is grown in arctic, humid regions and in some tropical highlands. Major world producers of wheat crops are China, India, Russia, the USA (Steduto et al., 2012). Wheat plants are characterized by diverse growing conditions because of the variability of soil types and crop management. For instance, winter wheat requires a cold period (vernalization) during early growth and it is photoperiod sensitive (plant response to the length of light and dark periods)

Vernalization is the ability of plants to flower as a result of exposure to a prolonged period of low temperature. It is an environmental stimulus that ensures that flowering occurs in the appropriate season of the year Dennis & Peacock (2009); Kim et al. (2009). Vernalization accelerates the progression from the vegetative phase (period between germination and flowering) to the reproductive stage where the reproductive organs are developed in many plants. Many plants grown in temperate climates require vernalization and must experience a period of low winter temperature to start the flowering process. Besides winter wheat, winter barley and rapeseed are other plants with vernalization requirements before flowering (Oliver et al., 2009; Ream et al., 2014).

Temperature is one of the factors determining growth and development of wheat plants (Porter & Gawith, 1999). The minimum mean daily temperature required for the growth is about $5°C$ while the mean daily temperature for opti-

mum growth varies between $15 - 23°C$. Recent studies from Chen et al. (2014) observed that an increase in mean night-time temperature by $1.1°C$ will enhance wheat development. Similarly, Benlloch-Gonzalez et al. (2014) demonstrated the reduction of the positive effect of elevated CO_2 on wheat root growth as a result of rising temperature. Higher temperatures can also catalyse senescence that could cause a significant reduction in yield.

Wheat can encounter moisture stress, the severity of which reduces with the developmental stage. On the other hand, excess water can also cause damage to the plants, for instance, being waterlogged during vegetative growth can reduce the wheat yield substantially. Wheat plants are also irrigated in some areas, especially in arid, semi-arid and Mediterranean regions. Irrigation is necessary to prevent water stress in these regions. Continuous supplies of water to wheat plants during the flowering stage is essential to maintain quality yield with high protein content (Passioura & Angus, 2010). Wheat is an important cereal crop, being the third largest crop grown in the world. Other cereal crops are discussed briefly below.

Barley is another cereal crop. It is adapted to a wide range of environmental conditions and it is commonly planted in temperate regions as a summer crop and in tropical regions as a winter crop. Barley is closely related to wheat except that it can withstand stressful condition much better than wheat. Barley is used mainly in beverage production and as animal feeds. Major world producers of barley are the Russian Federation, Ukraine, France, Germany, Spain and Australia. Rye is another cereal crop similar to barley and wheat. Rye is cultivated across Eastern, Central and Northern Europe, as well as USA and Canada. Oats are planted in temperate regions and are used as oatmeal and livestock feeds.

2.3.2 Rice

Rice is an important staple food worldwide. It is the main source of energy for more than half of the world's human population (Lampe, 1995). It provides about

20% of the calorie consumption of the world. It is planted all over the world but is more predominant in Asian countries with production capacity of about 523MT (Dubey, 2001). It is cultivated mainly between latitudes 50°N and 35° S. Water availability determines the distribution and cultivation of rice to some extent. Rice is extremely sensitive to water shortage. Different varieties of rice exist with varying degrees of water requirements. There are irrigated and rainfed cultivated rice crops.

The amount of cultivated cropland area for rice increases yearly. Rice is planted throughout the year and anaerobic decomposition of organic material in flooded rice fields generates CH_4 emission. Matthews et al. (1991) described the geographical and seasonal distribution of cultivated areas and emissions of rice. Rice responds positively to an increase in CO_2 concentration because it is a C_3 plant with an enhanced photosynthetic rate. C_3 plants use rubisco, an enzyme that can fix CO_2 to make a three-carbon compound, as the first stable product of carbon fixation. Over 95% of earth's plant species are C_3 plants, including temperate cereals. Rice is affected by temperature and temperatures above and below its optimal level tend to reduce the growth rate (Narciso & Hossain, 2002).

Demand for rice is predicted to increase because of the projected increase in population. At the same time, the cropland area for rice cultivation is predicted to reduce as water availability is expected to reduce. Therefore, it is important to increase the yield and productivity of rice cultivation in order to adequately feed the increasing global population (Gopalakrishnan et al., 2014). Shrestha et al. (2014) examined the responses of winter and summer rice yield under future climate change and investigated various climatic conditions that would negatively affect the rice cultivation, and some adaptation measures to overcome these problems were highlighted. Major rice producing countries are China, India, Indonesia, Bangladesh, Vietnam, Myanmar, Thailand and the Philippines.

2.3.3 Maize

Maize is the most widely planted cereal after wheat and rice. It was first cultivated in Central America but it is now grown throughout the tropics, especially in areas with adequate rainfall. There are many varieties of maize crop. Maize has many uses such as livestock feed, food, and in industrial biofuels. It is a common staple food in many tropical countries. It is cultivated on a wide diversity of soil, climate and provides about 36% of the global grain production. Maize is a C_4 plant and has high water-use efficiency. C_4 plants, unlike C_3, have biochemical and anatomical CO_2 concentration mechanisms that increase the intercellular CO_2 level at the site of fixation. As a result, they have a great reduction in carbon losses due to photorespiration which is a process that begins when rubisco fixes molecular oxygen, as opposed to carbon dioxide, which ultimately leads to the evolution of CO_2 from plants. C_4 plants are very common in tropical regions. Other examples of C_4 plants are sugarcane, sorghum and millets.

2.3.4 Oil crop

The main oil crops are soybean, rapeseed, groundnut and sunflower. Soybean has the largest oil cropland area globally and it is an important source of protein (Krantgartner et al., 2011). Sunflower is used as livestock forage and as a domestic cooking ingredient. The top producers of sunflower are the EU, Argentina, Russia and Ukraine. Heat is a major factor determining the growth rate of sunflower plants.

2.4 Representative Concentration Pathways (RCPs)

The RCPs (shown in Figure 2.2) are a set of four current and future concentration pathways for greenhouse gas developed for climate modelling as a basis for long-term and near-term modelling experiments. They are provided as input for modelling climate and as a basis for the assessment of possible climate

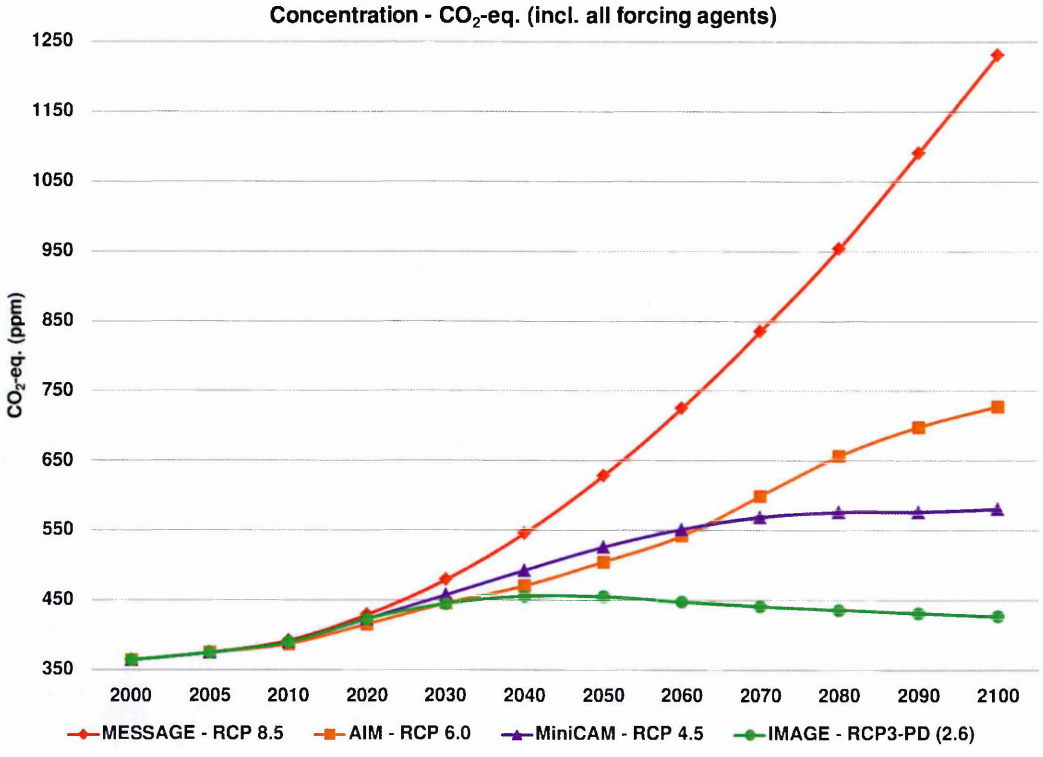


Figure 2.2: RCP scenarios: Note "eq" denotes equivalent

impacts and mitigation options. They are the four scenarios that are considered in this thesis. They describe alternative possible future emission scenarios and are designed to standardize climate model simulations for inter-comparison. They replaced earlier scenario-based projections of atmospheric composition, eg the Special Report on Emissions Scenarios (SRES) with details in Nakicenovic & Swart (2000). The RCPs can also be extended to 2300 called Extended Concentration Pathways (ECPs) (van Vuuren et al., 2011; Meinshausen, 2011a).

The pathways are characterised by either long-term stabilisation within the period $2100 - 2150$ (RCPs 4.5 and 6.0), progressively increasing forcing up to 2100 (RCP 8.5), or decreasing forcing by 2100 and beyond (RCP2.6, also named RCP-3PD). The four RCPs scenarios cover a mitigation pathway in which radiative forcing is reduced to 2.6 Wm^2 (RCP 2.6) by 2100, a business as usual pathway in which radiative forcing increases to 8.5 Wm^2 for RCP 8.5 by 2100, and two stabilization pathways in which, by 2100, forcing levels out at 4.5 Wm^2 for RCP 4.5 and 6.0 Wm^2 for RCP 6.0 respectively. The 5^{th} Assessment Report

(AR5) of the IPCC reports, RCP 2.6, engenders a world with global mean surface temperature stabilized at $1°C$ by the 2050s. Similarly, RCP 8.5 would lead to a global mean warming exceeding $1.4°C$ and up to $4.8°C$ by the 2080s (Moss et al., 2010; van Vuuren et al., 2011; Stocker, 2013).

Chapter 3

Literature review

This chapter reviews the literature on methods of modelling climate, the biosphere and crop yields, before reviewing relevant emulation methods. Works on climate change and terrestrial biosphere are reviewed in Section 3.1. Section 3.2 is focused on carbon fluxes while 3.3 focuses on crop yields before discussing various emulation strategies in Sections 3.4 and 3.5, 3.6. Both censored and Bayesian regressions are included in that discussion.

3.1 Climate change and biosphere

Numerous studies have investigated climate change and its adverse effects on terrestrial ecosystems (Heyder et al., 2011; King et al., 1997; Bergengren et al., 2011). Climate change causes an increase in atmospheric and ocean temperatures, a rise in sea level, an increase in the length of growing seasons and changes in ice-free periods and precipitation patterns. The adverse effects also include changes in the frequency and severity of extreme events, like severe storms and floods, which could affect the whole ecosystem.

Anthropogenic emissions of CO_2 and other greenhouse gases have been attributed as major external forcing on climate change. CO_2 absorbs long wavelength radiation, causing a rise in surface temperature i.e global warming (Kirschbaum, 2000; Cao et al., 2009; Shakun et al., 2012). A doubling of atmospheric CO_2 is

expected to increase the global temperature by (1.5-4.5) $^{\circ}C$ (Stocker, 2013).

The climate system plays a vital role in determining natural ecosystems. An ecosystem involves interactions between organisms and their environment. It is an interdependent system of plants, animals, and micro-organisms that interact with one another and the environment. Ecosystems are linked through food, water and energy transfer (Chapin et al., 2002; Kennedy et al., 2001). Elevated CO_2 causes increases in the photosynthesis rate of C_3 plants (Bazzaz, 1990). Most studies indicate an increase in photosynthesis rate as a result of elevated CO_2. In plants, the CO_2 effect is more pronounced when there is adequate water and nutrient availability. Some studies have also shown that a reduction in stomatal conductance will cause a decline in transpiration rate as a consequence of elevated CO_2 concentration, thereby leading to an increase in carbon productivity and biomass accumulation (Bazzaz, 1990, 1984; Brown et al., 1986).

Some recent studies have found that climate change has impacts on the nutritional quality and safety of food and that cereal crops planted under elevated CO_2 show a decrease in protein and micronutrients (Vermeulen, 2014). For the major crops (wheat, rice, and maize) in tropical and temperate regions, climate change without adaptation is projected to negatively impact production for local temperature increases of 2° C or more above late-20^{th}-century levels, although individual locations may benefit (medium confidence) (Stocker, 2013).

On a different note, some authors have implemented a process-based technique for global assessment. In particular, Melilo et al. (1993) estimated the NPP patterns across the world using current climate and CO_2 concentration through a process-based terrestrial ecosystem model (TEM). TEM is a simulation model that can evaluate the amount of carbon, nitrogen and some important fluxes using information on climate, soil, vegetation as an input to the model. They observed an increase of 16.3% in global NPP if CO_2 were to double with no climate change. However, with climate change and no change in CO_2 concentration, there was little impact on the global NPP. Finally, they concluded that more than 50%

of the global annual NPP is found in the tropical climate.

Closely related is the work of Berthelot et al. (2002), which also investigated the responsiveness of the terrestrial biosphere to changes in climate using the Institut Pierre Simon Laplace (IPSL) coupled GCM model with a global carbon cycle, which is forced by CO_2 emissions. They first described the terrestrial biosphere and carbon-climate simulation models. They observed a positive climate feedback on CO_2 as a consequence of a much higher negative climate impact on the land carbon uptake in the tropics than the positive impact observed in the sub-tropics and higher latitudes. A similar effect was observed by (Dufresne et al., 2002). Moorcroft (2006) studied two different factors of below-ground acclimation and compositional change and the degree to which they can change the temperature and moisture dependencies of decomposition rate. This could consequently affect the expected terrestrial ecosystem feedback to the atmosphere. He also suggested a more reliable way of predicting terrestrial biosphere responses by connecting terrestrial biospheric models with several empirical ecosystem measurements.

Bergengren et al. (2011) developed two new ecosensitivity metrics to examine the responsiveness of the plant community changes for the next 30 years. The work used an equilibrium vegetation ecology model (EVE) which applies ecological principles to establish a relationship between monthly mean climate and an equilibrium prediction of vegetation structure. Ten different climate simulations from the IPCC Fourth assessment report were used in the analysis. The findings indicated that plants are vulnerable to catastrophic damage as a result of climate change.

3.2 Carbon fluxes

This section focuses on work directly related to the study of NPP, FC and HR.

3.2.1 NPP and climate change

Process-based model simulations of terrestrial ecosystem models provide a useful tool for studying ecosystem processes, especially at multiple spatial and temporal scales, so as to investigate carbon fluxes (eg NPP) in response to climatic change (Tian et al., 1998, 2010). Empirical modelling that uses correlation between ecosystem and climate variables offers a means of projecting the climatic impact on ecosystems, but the use of process-based modelling to assess the global impact can complement the empirical approach (Melilo et al., 1993; Potter, 1993; McGuire et al., 2001). Some notable process based simulation studies are described in detail below.

Cao et al. (1998) considered a Carbon Exchange between Vegetation, Soil and the Atmosphere (CEVSA) model. This is a process-based model that predicts the NPP, soil organic carbon and net ecosystem production under current and future climate scenarios. The process-based model combined the processes entailed in the terrestrial carbon exchange between vegetation, soil and the atmosphere through photosynthesis and respiration. CEVSA integrates the biophysical, vegetation and biogeochemical models, as do other process-based models in the works of Melilo et al. (1993); Berthelot et al. (2002); Moorcroft (2006) and Bergengren et al. (2011). In addition, assessment of the correlation between NPP and climate variables was performed.

The study of Cao et al. (1998) indicated that precipitation is the most important factor determining the spatial distribution of NPP globally, having a correlation ρ of 0.58 between NPP and precipitation, and a low correlation $\rho = 0.20$ between NPP and temperature. They also observed a very low correlation between NPP and each of net radiation and relative humidity (Cao et al., 1998). Lastly, the CEVSA model projected a global NPP increase of 14.5 $GtCy^{-1}$ for a doubling of atmospheric CO_2 without corresponding change in climate, while a reduction of 2.2 $GtCy^{-1}$ of global NPP as a result of climate change with CO_2 concentration held constant. Interactions between elevated CO_2 and climate change

produced an increase in predicted global NPP of 12.6 $GtCy^{-1}$.

Churkina et al. (1999) considered the relative importance of water availability as a major limiting factor of NPP. They investigated different methods that have been used by various authors to introduce water budget limitation on NPP. In addition, they also estimated the correlation between NPP and water balance. Closely related to the work of Prince (1991), Schloss et al. (1999) compared the performance of 15 different terrestrial ecosystem models for simulating NPP. In that study, sensitivity of annual monthly NPP estimates to seasonal climate variables and Normalized Difference Vegetation Index were performed. NDVI is an indicator that provides a measure of estimating photosynthetic capacity of vegetation canopies. Some of the models are run with NDVI instead of climate variables. Rank correlations between NPP and these climate variables (precipitation, temperature, solar radiation) and to NDVI were obtained. They observed large sensitivities in regions where NPP is limited by both temperature and precipitation.

Bondeau et al. (1999) presented several analyses on how differences in vegetation canopy and its phenology affect simulated NPP seasonally. They also compared the seasonal NPP estimate from the model with seasonal satellite data using the fraction of photosynthetically active radiation absorbed by the canopy. This is similar to the method used by Prince (1991). They concluded that seasonal changes in canopy largely determine the seasonal estimate of net primary productivity.

Peng et al. (1999) examined trends in the NPP of forests, focusing on Central Canada. They also used a process-based ecosystem model, CENTURY, similar to the method of Melilo et al. (1993). They investigated the sensitivity of NPP to climate change, CO_2 concentration and the frequency of fire disturbance. They observed an increase in NPP for both a change in climate and an increase in CO_2, as well as an increase in the incidence of fire disturbance. They attributed the increase in NPP to ecosystem feedback through an increase in net nitrogen

mineralization. An increase in decomposition rate causes an increase in nutrient availability to plants.

Cramer et al. (2001) predicted the impact of climate change and increasing CO_2 levels on the terrestrial carbon cycle. They compared the performance of six dynamic global vegetation models forced with a transient increase of CO_2 emission, similar to Melilo et al. (1993) and King et al. (1997). Similarly, Berthelot et al. (2002) investigated the terrestrial biospheric response to change in climate using a coupled model with a global carbon cycle. The model was forced by CO_2 emission. They observed a positive climate feedback on CO_2 similar to Dufresne et al. (2002). Nemani et al. (2003) reported the responses of global vegetation to climatic changes using both climatic and vegetation data. A biome-specific production efficiency technique was used to predict the annual and monthly net primary productivity. They observed that nitrogen, forest regrowth, climate and CO_2 fertilization are the major factors responsible for the increase in NPP in the Northern mid-latitudes, high latitudes and in a tropical climate.

In contrast to the process-based method of Peng et al. (1999); Berthelot et al. (2002); Moorcroft (2006); Bergengren et al. (2011), the earlier work of Prince (1991) estimated NPP from remote sensing data. He used a linear relationship between absorption of photosynthetically active radiation (APAR) and Normalized Difference Vegetation Index (NDVI) to predict net production and respiration. Similarly, Imhoff et al. (2004) also used the combination of satellite data and a terrestrial carbon model to assess the effect of urbanization on NPP in the United States. They observed a significant negative impact of urbanization on NPP. Several studies emphasised the role of precipitation and water availability as a major driver of NPP distribution (Cao et al., 1998; Churkina et al., 1999; Lieth, 1975; Neilson et al., 1994; Rosenzweig, 1968; Weltzin et al., 2003).

On the other hand, Heyder et al. (2011) analysed the global impact risk of terrestrial ecosystem change due to climate change. They used a newly derived impact metric which estimates the distance of a possible future ecosystem change

from current conditions using vegetation structural change and biogeochemical shifts. The metric was based on the change in the biogeochemical flux of carbon, water, and carbon storage.

We can easily infer from these studies that NPP is affected by climate change and there is a need for further study in this area to broaden the scope of the impact assessment. Two of the aims in this thesis are: Firstly, to assess the response of global NPP flux under the interactions of elevated CO_2 and future climate change, and secondly to find out if statistical emulation can provide a better alternative for studying the interactions of global NPP and climate instead of a process-based technique. The next section looks at the past studies on the interactions between heterotrophic respiration and climate change.

3.2.2 Heterotrophic respiration (HR) and climate change

Lloyd & Taylor (1994) examined the effect of temperature on heterotrophic respiration. Osborne & Wheeler (2013) studied the impact of soil moisture on the heterotrophic respiration. Temperature and moisture strongly influence the heterotrophic respiration (Ise & Moorcroft, 2006). Similarly, Smith et al. (2013) examined the sensitivity of heterotrophic respiration to temperature. In another related study, van Wart et al. (2013) investigated how global warming and changes in precipitation patterns could affect the rate and temperature sensitivity of heterotrophic respiration. The early work of Cook & Orchard (2008) observed a linear relationship between soil respiration and water content. Moyano et al. (2012) fitted a linear model to also establish a relationship between respiration and moisture using various soil properties. They first applied a generalised additive model to obtain a smooth curve to the proportional response (PR) of soil respiration to moisture. A linear regression was then used to determine the relationship between PR and soil properties (soil pore space, bulk density, soil organic carbon and sand).

Maxwell et al. (2013) predicted the soil respiration using variables like temper-

ature, moisture and water table depth. They considered six different methods of predicting soil respiration from peatlands. Four of these methods are documented in the literatures (Ricker, 2011; Tuomi et al., 2008; Horion et al., 2013; Lloyd & Taylor, 1994). They proposed variations of the technique used by (Horion et al., 2013). This approach incorporated water table depth as an additional variable that affects soil respiration.

Suseela et al. (2012) investigated the joint impact of global warming and change in precipitation pattern on heterotrophic respiration at ecosystem level. A mixed model was first used to assess the main and interactive effects of warming and precipitation change on respiration. They later analysed the temperature sensitivity of heterotrophic respiration (HR). In this study, HR was fitted as an exponential function of temperature, similar to Tuomi et al. (2008). They further fitted HR as a quadratic of both soil temperature and water content in order to assess the interaction between moisture and temperature on respiration. Their conclusion was that global warming accelerates HR only during spring and that HR is independent of warming during the summer and fall.

Tuomi et al. (2008) applied six different models to investigate the sensitivity of soil heterotrophic respiration to temperature. The models include simple exponential functions, a Gaussian model and a very complex Del Grosso model. They also considered an extension of an exponential model that combined an exponential function and some additive parameters. The performance of each model was compared using their residual sum of squares and Bayesian Information Criterion (BIC). The results indicated that both the Gaussian model and the exponential function with some additive parameters clearly reproduce the pattern of decreasing Q_{10} with an increase in temperature. Q_{10} is the relative increase of the respiration rate when temperature increases by 10°C. However, a mixture of an exponential function with some additive parameters only provided good estimates at high temperatures.

Similarly, Smith et al. (2013) examined, at the ecosystem level, the sensitivity

of heterotrophic respiration to temperature. They analysed the intrinsic temperature sensitivities of ecosystem respiration for different climates. Ecosystem level sensitivity of $Q_{10} < 2$ was observed. Apart from this work, some notable studies have also documented that climate warming increases the rate of heterotrophic respiration, thereby causing a positive feedback effect to increasing CO_2 levels in the atmosphere (Wang et al., 2014; Schindlbacher et al., 2009).

In a related study, autotrophic respiration is another significant carbon flux that affects ecosystem carbon balance, and the net ecosystem carbon flux often changes as the balance between photosynthesis and plant respiration changes (Ryan, 1991). Autotrophic respiration offsets more than 50% of the carbon fixed in photosynthesis and may regulate productivity and carbon storage in forest ecosystems (Ryan et al., 1997). The respiration rate for any particular plant increases exponentially with ambient temperature (Ryan, 1991; Turnbull et al., 2001). Autotrophic respiration is one of the primary outputs of LPJmL.

Katterer et al. (1998) examined the functional relationship between decomposition rates and temperature based on a review of published data. The data were analysed using two different dynamic models to fit a least squares regression to the incubation temperature and carbon evolution. Reich & Schlesinger (1992) analysed the dependence of temperature on the rate of decomposition from CO_2 measurements from different soils. They used various data from different ecosystems and expressed observed carbon fluxes as a function of mean annual temperature.

Carter et al. (1998) observed a decline in soil organic carbon and nitrogen with an increase in temperature. Holland et al. (2000) investigated the sensitivity of heterotrophic respiration to temperature changes with different laboratory experiments, by examining the variability of temperature regulation on heterotrophic respiration among different vegetation and soil types. They assessed the consequences of temperature variability on ecosystem modelling using the CENTURY terrestrial ecosystem model that has the potential of analysing the impact of

temperature regulation on decomposition and heterotrophic respiration of carbon and nitrogen fluxes. An exponential increase in tropical soil CO_2 with increasing temperature was observed under optimal soil conditions. However, a decreasing response was observed under limited substrate availability, and there was a high correlation between heterotrophic respiration and various soil organic carbon and nitrogen content. Overall, soil carbon is more sensitive to temperature regulation than soil nitrogen content and net nitrogen mineralization.

In contrast to the work of Carter et al. (1998), Giardina et al. (2000) provided evidence suggesting that the decomposition rates of organic compounds do not change with temperature for some forest soils but are fairly constant on a global scale. They inferred that it was not only temperature changes that are responsible for the decay of forest carbon. Braswell et al. (1997) examined the response of terrestrial carbon storage to climate variability using CO_2, temperature and vegetation index data derived from satellite information. They carried out three different sets of analysis and found that ecosystems exhibit a complex behaviour in response to climate variability. One of the analyses involved regression of NDVI and temperature anomalies using a weighted least square regression to account for large correlation. The results indicated a significant correlation between these two variables that varies across different ecosystems. They concluded that there was diversity in the biospheric response of the ecosystem to climate change.

Moreover, Lloyd & Taylor (1994) reassessed the published data to investigate the suitability of linear, exponential and Arrhenius models earlier used in the work of Reich & Schlesinger (1992) and Braswell et al. (1997) for predicting the relationship between soil respiration and temperature. The implication of each individual method for the seasonal cycle of soil respiration and net ecosystem production was further examined. They concluded that additional information is required on temperature sensitivity of carbon input to the soil through NPP to better understand the effect of temperature changes on soil carbon.

Kirschbaum (2000) extended the work of Lloyd & Taylor (1994) by investigat-

ing the relative impact of temperature changes on the decomposition rates of soil organic carbon and NPP. In addition, he also examined whether changes in soil organic carbon would act as a positive or negative feedback on climate change. He observed that the decomposition rate of organic carbon is more sensitive to temperature than for NPP, and as a result, could potentially cause a significant positive feedback from soil organic carbon into the atmosphere. Nevertheless, he concluded that the overall feedback will be relatively small, because the positive feedback observed could be easily offset by the beneficial effect of CO_2 fertilization, which is capable of enhancing plant productivity and soil organic carbon storage.

3.3 Modelling crop yields

This section focuses on a general review of crop yield modelling. The following notable agricultural crops, namely temperate cereals, rice, maize and oil crops (sunflower, soybeans, rapeseed, groundnut), are considered in this thesis. The relationships between crop yields, weather and climate have attracted considerable attention. Weather is the state of atmosphere (temperature, precipitation, cloudiness) while climate is the accumulation of daily and seasonal weather events over a long period of time. Impacts of climate change on crop yields are already happening across the world and adverse impacts are common globally. However, in some places, for instance in the UK and China, positive impacts of climate change on crop yields are observed (Stocker, 2013; Vermeulen, 2014).

The world population is projected to increase by 35% by the middle of this century (Crossette, 2010). This will cause a rise in demand for major food crops that will necessitate a considerable increase in crop production. Climate change, food insecurity and how to effectively feed over 9 billion people by mid-century are major problems threatening human prosperity (Dobermann & Nelson, 2013; Godfray et al., 2010; Smith et al., 2013). Rising temperatures and CO_2 levels, with the change in precipitation patterns, are important biophysical effects that

will affect crop production (Parry et al., 2004). The problems highlighted above require multi-disciplinary approaches and, to effectively tackle them, a robust and coherent assessment of the climatic impact on future crop-yields is essential to inform policy makers. Adequate knowledge of crop yield responses under various climatic scenarios is essential for agricultural policy implementation and global food security.

Kart (1979) examined the empirical relationship between crop and weather using a ridge regression approach. He used empirical relations between crop and weather for the yield predictions. He observed that most of the estimated parameters are not stable and that adequate care must be made in choosing statistical methods for developing models for crop yields. Miller (1976) analysed similar data using a least squares regression method. Reddy et al. (2000) estimated changes in crop yield from monthly weather projections of climate variables. Similarly, Wallach (2011) extended the Kart (1979) approach by estimating wheat production using a multiple regression approach, this involves the concept of using a weighted amount of rainfall rather than actual rainfall and this was found to improve the predictive power of the model. In a related analysis, Bornn & Zidek (2012) evaluated wheat prediction using a Bayesian method. They examined the significance of incorporating spatial information in crop-yield modelling and the consequence of neglecting such important information.

Drummond et al. (2003) compared the performances of several statistical techniques for modelling the relationship between grain yield and soil properties. They adopted linear and non-linear methods and evaluated them. They further compared the predictive performance of a stepwise multiple linear regression. Schlenker et al. (2006) investigated the effect of change in average weather on crop-yield, focusing more on the non-linear effect of temperature on growing season. However, this study did not incorporate CO_2 fertilization effects, which is the crop response to elevated CO_2 emission in the assessment, and which can substantially increase plant growth.

In contrast to the regression approaches of Wallach (2011), Schlenker et al. (2006), Osborne et al. (2013), Matis et al. (1989) and Jain & Agrawal (1992) several researches applied a non-parametric Markov chain approach to crop yield prediction. Kim et al. (2005) applied a Bayesian bootstrap method to similar data by obtaining the posterior distribution of the parameters rather than the distribution of a sample statistic. Sacks et al. (2010) described the relationships between climate and the dates of planting and harvesting for some notable crops.

A phenological approach to examine how vegetation dynamics is dependent on climate variability is discussed in Horion et al. (2013). In a related study, Osborne & Wheeler (2013) examined crop yield variability through time and assessed whether such variability can be attributed to climate change. In the same vein, Lobell (2013) investigated the impact of measurements error on statistical models. He focused on the effect of inherent error from climate ensembles on the quality of statistical crop models. van Wart et al. (2013) performed yield gap analysis relative to different agro-climatic zones. Interactions between climatic cycles and crop productivity were investigated by Maxwell et al. (2013), while Liu et al. (2013) evaluated crop yield predictions under various soil tillage systems.

Similarly, Kitchen et al. (1999) compared the predictive performance of stepwise multiple linear regression with that of projection pursuit regression and neural networks. These methods were used to model the relationship between yield and soil properties. It was found that neural networks performed better than the other methods. In related studies Schlenker et al. (2006) and Sacks et al. (2010) considered multiple linear regression techniques using polynomial and interaction terms for modelling crop productivity measure.

In contrast to the least squares approach, Lobell et al. (2006) developed a non-linear model to relate weather and climate to crop yields. Chebyschev coefficients derived for the probability distribution of temperature were used. They observed a significant nonlinear relationship between temperature (degree days) and yields. Osborne et al. (2013) described the relationships between planting and harvesting

dates with climate for some notable crops. Several studies in the literature have explored the impact of climate change on crop yield. However, very few studies have examined the combined effect of climate and CO_2 fertilization on crop yields under various emission scenarios and GCMs. For instance, Parry et al. (2004) used only one GCM but in combination with inclusion of CO_2 fertilization, while Sacks et al. (2010) incorporated several GCMs in their analysis but did not assess the impact beyond 2030. Osborne et al. (2013) also used several GCMs but focused only on the single time point of 2050s and one emission scenario.

Not many studies have evaluated the impact of crop-yield at a global scale; most have concentrated on smaller regions within a country. For instance, Reddy et al. (2000); Schlenker et al. (2006) and Drummond et al. (2003) only focused on the United States, Bornn & Zidek (2012) concentrated on a smaller region consisting of a Canadian Province, while Osborne et al. (2013) focused on the 30 top producers of wheat and soybean and Liu et al. (2013) focused on Northeast China.

Consequently, none of the above literature adequately examined the joint impact of climate change and CO_2 fertilization effects under various crop management options, emission scenarios and GCMs on a global scale. Consequently, very little is known about how crop-yield will respond to climate change under different levels of management practice combined with CO_2 effects and emission scenarios. Hence further research is needed in this area.

The work reported in this thesis will extend existing works in this area by emulating the joint response of potential crop-yield to climate change under various future scenarios, simultaneously incorporating several Representative Concentration Pathways (RCPs), General Circulation Models (GCMs), CO_2 fertilization effects, and crop management practices. GCMs are numerical models representing the climate systems while RCPs are the CO_2 emission scenarios adopted by IPCC for the running of global climate models. The CO_2 fertilization effect is the crop response to elevated CO_2 emission.

3.4 Emulation techniques

Several methods have been used to construct statistical emulators for modelling purposes (Santner et al., 2003). There are Monte Carlo approaches that give reliable method when the simulation code is fast to run (Higdon et al., 2003; Kettleborough et al., 2007). Gaussian process emulators are a more efficient approach but have limited ability for modelling high-dimensional or time-dependent analysis. (Currin et al., 1991; Kennedy et al., 2001; Oakley & O'Hagan, 2002).

A filtering method is applicable when simulation data is in a sequential form, while an ensemble Kalman filter can additionally handle multivariate parameter estimation (Annan et al., 2005). Numerical integration has been applied to low dimensional problems and MCMC methods have also been used, although these involve many iterations to reach convergence (Annan et al., 2005). Annan et al. (2005) used an ensemble Kalman filter for parameter estimation and forecasting in climate modelling. Kalman filtering is often applicable for integrating noisy measurements. A very fast simulated annealing method was used in Jackson et al. (2004). Spline interpolation has been used for equally spaced data points by constructing polynomials of low degree in a regression method (Faraway, 2006).

Shahsavani et al. (2011) used a sequential adaptive design in combination with polynomial regression to develop a surrogate model for the estimation of sensitivity indices for different sets of inputs. The method can be applied when there is no prior information on the response surface and the objective is to examine the global variability in the model. A dominant mode analysis was discussed in Young (1999) and the method was used as an emulator to extract the dominant mode of a higher order dynamic model. Young et al. (2011) described the behaviour of large linear dynamic models using the statistical principle of dynamic emulation. The approach identifies a low-order dynamic model that could approximate the behaviour of the higher-order dynamic simulator at a low computational cost.

Emulation can also provide a measure of the uncertainties associated with the

projections. There are many sources of uncertainty in projected climate changes. For instance, there may be deficiencies in modelling key processes that regulate important biophysical effects, such as the water and carbon cycles. Also, boundary conditions for different global climate models can introduce uncertainty, as can regional climate variability (Christensen et al., 2007). Nevertheless, generating ensembles of simulations can provide a useful means of quantifying the uncertainty in projections of regional climate changes (Graham et al., 2007; Beniston et al., 2007).

Addressing a different issue, a number of works have mainly focused on uncertainty quantification. For instance, Osborne et al. (2013) quantifies the uncertainty introduced by different GCMs in a crop impact study. They observed that the impact of climate change on crop yield is mostly attributed to the change in growing season temperature and that even under a policy of no adaptation the degree of impact varies both from one crop to another crop and between countries.

Webster & Sokolov (2000) and Kettleborough et al. (2007) quantified uncertainty in a climate model using the combination of a deterministic equivalent modelling technique and a Monte Carlo approach. Their assumption was that, if the model response can be represented by a polynomial, then the response can be easily approximated by deterministic equivalent models. Webster et al. (2003) examined uncertainty in emission projections under different policy scenarios. Recently, Osborne et al. (2013) quantified the uncertainty associated with a global climate model (GCM) in a crop impact study, confining their attention to soybean and spring wheat.

3.5 Censored regression

This section discusses literature on existing methods for censored data. Censored regression is performed when the variable of interest is not observed over its entire range; therefore it often causes estimated mean and variance to be biased. For instance, in a study following patients who have been treated for cancer, the

study will stop after a set period of time, and the survival time of patients who survive a long time will not be known.

Censored regression is closely related to truncated regression models that arise when there are missing observations in both dependent and independent variables. For instance, let $Y = [y_1, \ldots, y_n]$ be censored data when we observe $X = [x_1, \ldots, x_n]$ for all observations, but we only know the true value of Y for a limited range of observations. The values of Y in some ranges are reported as a single value, such as zero or clustered around zero (Schnedler, 2005). Table 3.1 shows different types of censored data as left, right, interval censored data and a complete observation. With left-censored data we do not know the precise value of small observations; with right-censored the precise value of the large observations are unknown; and with interval censoring, the data are known to lie in one of a number of intervals, but the precise values are unknown.

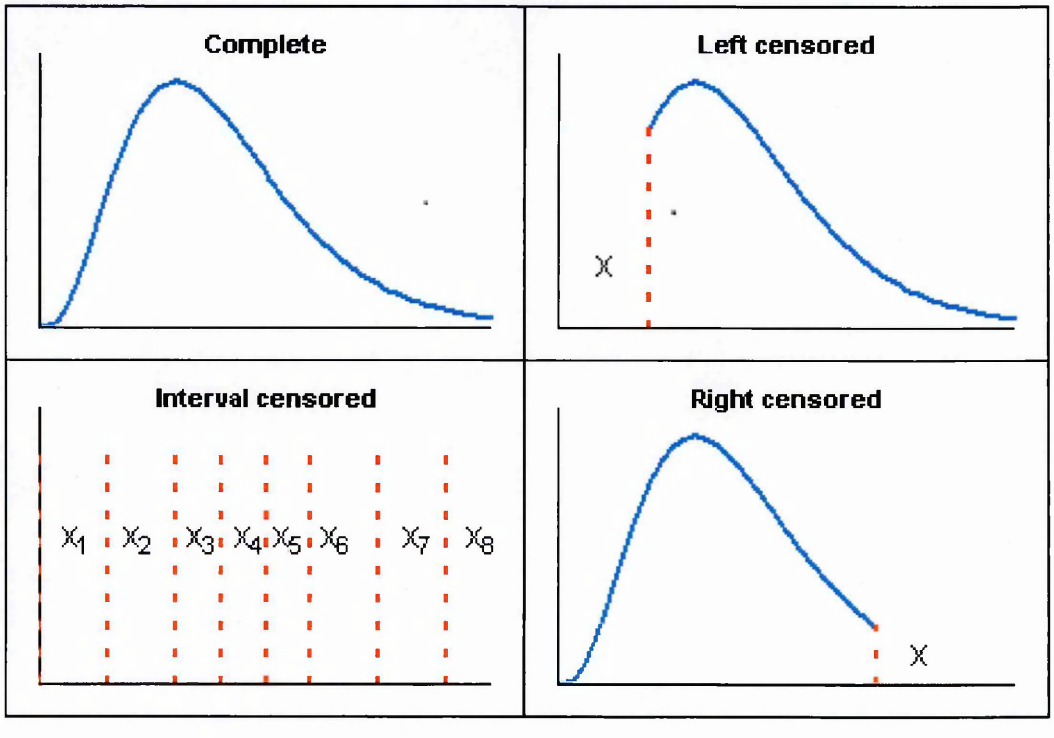


Figure 3.1: Diagram illustrating the distribution of different types of censored data.

Commenges (2002) applied multi state models to interval censored data. Multi-state models can be defined as any process that takes a finite number of states. These processes are usually denoted by their transition probabilities,

which in a Markov-chain representation are the probability of transitioning from one state to another. The motivation for applying this technique is based on the deficiency of information in the data. The data can only be observed at a discrete point in space. Early work of Amemiya (1973) estimated the regression parameters of a normally distributed dependent variable that is truncated to the left of zero. Schnedler (2005) described the maximum likelihood estimation for multi-dimensional censored data.

Miller (1976) analysed right censored data with a least squares regression method. The approach was based on the product-limit estimator of a distribution function of right-censored data. The Kaplan-Meier (product-limit estimator) is a non-parametric maximum likelihood estimator of a distribution function. It is an empirical method of estimating a survival function from right-censored observations. It relies on simple assumptions based on information from the censored samples in its estimation (Kaplan & Meier, 1958, 1992).

Zhang & Li (1996) used Buckley James Ritov-type estimators for the estimation of linear regression parameters of doubly censored data. Double censored data occurs when there is presence of both the left and right censored observations. The Buckley-James method, described in Buckley & James (1979), is an extension of linear regression estimators to censored variables. It is an iterative and updating procedure for estimating regression parameter from censored data. It is based on a non-parametric estimation of the residual distribution to deal with censoring. One major caveat of this method is the difficulty of computing the variance estimator of the Buckley-James estimator. Lai & Ying (1991) modified this approach slightly to accommodate asymptotic variance in order to avoid the difficulty associated with the variance estimation.

Closely related is the work of Potter (2000), which generalized the Buckley-James technique to a multivariate case based on a non-parametric method. Similarly, Ritov (1990) also applied this iterative technique on censored data, replacing unobserved Y values with their estimated conditional expectation. Miller &

Halpern (1982) compared the performance of four regression techniques applicable to censored data. These are the methods of Cox (1972) which implemented a partial likelihood estimation of regression parameter, and the methods of Miller (1976); Ritov (1990); Buckley & James (1979). Miller & Halpern (1982) concluded that Cox and Buckley & James estimators are more reliable than other methods, particularly for censored data. The other methods also have some methodological limitation so for instance, Miller's estimator is influenced by the censoring pattern.

Addressing a different issue, Jammalamadaka & Mangalam (2003) considered non-parametric maximum likelihood estimation (NPMLE) from middle-censored responses using the self-consistency equation. Middle censored data arise when the observations are not fully observed (incomplete) in some random interval. Further work on self-consistent estimators of interval censored data can be found in Chang & Yang (1987). Recently, Frydman & Liu (2013) also applied this non-parametric maximum likelihood procedure to the distribution function of interval censored data.

Kim et al. (2005) applied the Bayesian bootstrap method to doubly censored data while Rubin (1981) had earlier generated the posterior distribution of the parameters rather than sampling the distribution of a statistic. The posterior probability for the Bayesian bootstrap is computed from the Gibbs sampler algorithm, which is constructed from the empirical likelihood. This method is simpler and faster than the usual bootstrap algorithm.

Ren & Gu (1997) analysed doubly censored observations using M-estimators, which is a robust technique of parameter estimation especially in the presence of many outliers. Let $r_j = (Y_j - \hat{Y})$ be the residual from a fitted model, the least squares method minimizes the function $\sum(Y_j - \hat{Y})^2$. This minimization often results in inconsistent estimates if some observations are extreme. M-estimators can be applied to reduce the effect of distortion from the extreme observations by replacing the residual with a linear function, see further details in Ren & Gu

(1997); Shen (2013) and Szatmari-Voicu (2011). Ren (2003) applied a similar method but focused more on an independent identically distributed version of doubly censored data. In contrast, Zhao & Hu (2013) focus on right censored-data and use a non-parametric semi-Markov model where the transition intensities are a function of the present state and its duration. On a different note, Ren (2008) applied a weighted empirical likelihood-based method which is essentially a semi-parametric maximum likelihood estimator applied to double censored data. Lin et al. (2012) used quartile regression.

3.6 Bayesian emulation

This section provides a review of Bayesian techniques for emulation. Under a Bayesian perspective unknown parameters are treated as random variables. A Bayesian framework for emulation is almost always based on the assumption that a Gaussian process prior distribution can be specified for unknown parameters and hyperparameters. The given prior distribution can be updated from training data. Applying Bayes rule to this setting, a posterior distribution can be obtained. The posterior distribution is also a Gaussian process. The posterior distribution of the parameter that determine the function $f_j(.)$ in equation (2.1) is referred to as an emulator. Then $f_j(.)$ is a GP with mean and covariance functions $m(.)$ and $Cov(f_j(.), f_j(.'))$ respectively.

Tebaldi et al. (2005) proposed a Bayesian statistical model that combines information from a multi-model ensemble of Atmospheric Ocean General Circulation Models (AOGCMs) with observations in order to determine the probability distribution of future climate change. In the same vein, Lopez et al. (2006) developed a Bayesian statistical model to produce probabilistic projections of regional climate change using observations and ensembles of GCMs. A probabilistic method to overcome the challenge of inclusion of model inadequacy in order to improve the quality of a simulator ensemble was presented in Kennedy et al. (2008) and Rougier (2007). Gosling et al. (2007) estimated the uncertainty asso-

ciated with the carbon flux simulated from a dynamic global vegetation model. They used a GP emulation technique proposed by Oakley & O'Hagan (2002) to quantify these uncertainties for England and Wales.

Similarly, Higdon et al. (2008) applied the Oakley & O'Hagan (2002) approach in conjunction with a PCA for basis representations of high-dimensional output. Apart from reducing the dimensionality of the problem this PCA technique also reduces the computation time required for obtaining the posterior distributions. A closely related study of Wilkinson (2010b) performed a calibration of multivariate experiments by extending the approach of Kennedy et al. (2001) to multivariate models. This study incorporated principal component analysis, like Higdon et al. (2008), to project the multivariate output to a lower subspace. Boukouvalas et al. (2008) focuses on a variety of dimension reduction methods, like principal component analysis, independent component analysis and a Gaussian process latent variable model. They illustrated the methodologies by modelling time series output data.

Young et al. (2011) described the behaviour of large linear dynamic models using statistical principles of dynamic emulation. Their approach identifies a low-order dynamic model that approximates the behaviour of the high-order dynamic simulator at a low computational cost. A dominant mode analysis of Young (1999) is used to extract the dominant mode of a higher order dynamic model as an emulator for the construction of a low order dynamic model. In other words, it is the statistical identification and estimation of linear differential equation models.

Kennedy et al. (2001) applied a Bayesian technique to calibrate computer models. They obtained the posterior distribution of a parameter that measured residual uncertainty. Oakley & O'Hagan (2004) described a Bayesian method for quantification of uncertainty in complex computer models. They also extended variance decomposition and regression approaches to perform probabilistic sensitivity analysis under a Bayesian framework. This method performed better

than Monte-Carlo-based and Latin hypercube sampling methods. Kennedy et al. (2006) described some notable examples where GP modelling applications have been implemented.

A major difficulty with GP modelling is the computational effort associated with dealing with large data, as computer time scales are of order $O(n^3)$ where n is the number of observations. Several techniques have been adopted to overcome this computational problem. Earlier techniques are documented in Rasmussen & Williams (2006) and Quinonero-Candela & Rasmussen (2005). In particular, Shi et al. (2005a) used hierarchical Gaussian process mixtures for regression of a large data-set with repeated measurements. Also, Shi et al. (2003), focusing more on heterogeneous data, use a hybrid Markov chain Monte-Carlo method, while Rasmussen (1999) demonstrated a Markov Chain Monte Carlo (MCMC) implementation of a hierarchical infinite Gaussian mixture model that is similar to a Dirichlet process mixture model. Dirichlet process modelling is a method of assigning a high-dimensional probability distribution to categorical values and is common in Bayesian non-parametric models (Hannah et al., 2011). The method performs better for multidimensional data.

Subsets of regressors is another popular approach for approximating high dimensional data (Brailovsky, 1987; Poggio & Girosi, 1990). For large data sets, the standard GP technique involves inversion and the solving of large systems of linear equations. With large dataset the computation that is required is often too great to be practicable. Subsets of regressors provide a low rank approximation technique. For a subset of size p, the idea is to reduce the time taken to obtain the predictive mean from $O(n^2)$ to $O(p)$ and for the variance from $O(n^3)$ to $O(p^2)$. The method involves partitioning of the covariance matrix to a low order. Another closely related approach is to choose a subset of the dataset. Similarly, Luo & Wahba (1997) implemented a mixture of regression and smoothing spline to estimate multivariate functions. Reduced rank approximation of the Gram matrix that involves speeding up the inversion of the covariance matrix $K_{n\times n}$ was demon-

strated by (Bunea et al., 2011) and (Drineas & Mahoney, 2005b). They reduced the matrix dimension to a low rank p using the Sherman-Morrison-Woodbury formula. This gives the covariance matrix as $K_{p\times p}$ for $p < n$, see details of this method in Riedel (1992).

In the same vein, Sang & Huang (2012) extended the previous works of Bunea et al. (2011) and Drineas & Mahoney (2005b) by providing an approximation scheme that combined a reduced rank covariance method with a sparse matrix technique. Further modification to the residual of the reduced rank approximation was then applied. Recent works on reduced rank approximation can be found in Hirano (2014) and in the work of Solin & Sarkka (2014), who applied expansion of Laplace operators to eigendecompositions of covariance functions. An extension of the Skilling algorithm (Skilling, 1993), which is an iterative technique for decomposing large matrices, was discussed in Gibbs & MacKay (1997). The strategy is to limit the maximum order of matrix operation to $O(n^2)$, a procedure similar to the conjugate gradients method in Seeger (2000) and Brown et al. (1994).

A Bayesian Committee Machine (BCM) is a combination of ensembles of estimators on different datasets. In BCM, training data are sub-divided into smaller set of different sample sizes that are trained independently, but their prediction results are combined together with a covariance-based weighting scheme (Tresp, 2001, 2000). This is closely related to the Nystrom approach documented in Williams & Seeger (2001), which can be used for approximating a large set of eigenvalues and eigenvectors. A filtering approach discussed in Shi et al. (2005b) is a two-stage procedure, where a smaller filtered dataset is generated to represent the original data. Prediction is then based on the filtered data. The approach also involves a Nystrom approximation of eigendecompositions of the covariance matrix.

In order to examine the contribution of various inputs to the uncertainty in model output, Oakley & O'Hagan (2004) also illustrated a Bayesian technique

for probability sensitivity analysis that is computationally more efficient than MCMC methods. Svenson et al. (2014) extended Oakley & O'Hagan (2004) by using the procedure for sensitivity analysis based on a GP. They further derived quadrature-based methods of estimation with GP using Gaussian or Bohman correlation functions. The next section focuses on the multivariate technique as applied to GP emulation.

3.7 Multivariate GP emulation

Multivariate emulation deals directly with multiple outputs when diverse treatments have been applied to datasets. Some studies considered the outputs as a combination of separate univariate output, and emulated each independently. The principal disadvantage of this simple approach is that the data are assumed to be independent, which may not be true in general.

Some authors have emulated linear combinations of the output responses with a single emulator. A limitation of this approach is that it increases the computational burden into a two- stage method by first evaluating the best linear combination and then emulating the results. Another problem is that it can be difficult to back-transform the results to the original scale for proper interpretation. A related idea is to incorporate an output index as an additional input in the emulation, see Fricker et al. (2010) and Conti & O'Hagan (2010) for details of this approach.

Heitmann et al. (2006) applied a singular value decomposition to emulate multivariate responses and in the same vein Higdon et al. (2008) used a principal component approach to reduce the dimension of the response data. On the other hand, Bayarri et al. (2007) applied a wavelet technique to a set of functional outputs. Closely related are works of Morris et al. (2003) and Morris & Carroll (2006), where a Bayesian wavelet approach is also investigated. Similarly, Wilkinson (2010a) illustrated the application of Bayesian calibration to multivariate outputs and applied the PCA to project the response data onto a lower

dimensional space.

Fricker et al. (2010) illustrated both convolution and coregionalization methods for multivariate emulation where a non-separable covariance structure is implemented. A linear model of coregionalization utilizes eigendecompostion of the covariance matrix. A similar procedure is found in the works of Skilling (1993) and Shi et al. (2005b). The results of their studies clearly indicated the superiority of a non-separable covariance method over the separable covariance method. A common disadvantage of this approach is the problem associated with the inversion of the covariance function for a large dataset.

Rougier (2008) provides a comprehensive review of multivariate emulation techniques. He also illustrated an outer product emulation technique with a separable residual covariance function for multivariate emulation. The advantage of this approach is that it has been demonstrated to speed up the required time for emulator construction. Rougier et al. (2009) extended Rougier (2008), but enables the inclusion of expert opinion in the choice of parameter setting. Conti et al. (2009) demonstrated the dynamic emulation of multiple outputs. A more detailed treatment can be found in Conti & O'Hagan (2010), where an application to a dynamic vegetation model is given. The derivation of a multivariate extension of the Oakley (1999) technique is described in Hankin (2012).

3.8 Conclusion

Having reviewed the relevant literature on climate change and elevated CO_2 effects on terrestrial biosphere and various models that could be used to analyse such data, we can deduce that the terrestrial biosphere is indeed sensitive to climate change and CO_2 emissions. Emulation techniques reviewed in this thesis are applicable in diverse areas. We shall explore some of these methods that are more relevant to vegetation-climate impact studies, which is the aim of this thesis. Particularly, we shall consider the OLS, PCA, WLS, CR and GP methods on our crop yield and carbon fluxes data. We shall give full details of these methods

and their procedures in Chapter 5. Results and comparison of the performances of these methods will also be covered in Chapters 6, 7 and 8. These selected methods will enable us to assess the climatic impact on vegetation and provide a measure of uncertainty in the biospheric response to climate change.

Our data consists of simulated historical (1901-2000) and future (2001-2100) carbon fluxes and potential crop yields measurements at 0.5 by 0.5 degree spatial resolutions. Data analysis in this thesis will be demonstrated with the selected methods for the design of statistical emulators as a surrogate to LPJmL. Then we will compare the performance of each method in its ability to predict reliable estimates for various scenarios.

In the next Chapter we shall give full descriptions of impact models used for the simulation data in this thesis. The descriptions will comprise the climate (MAGICC6) and impact simulating models (LPJmL), as well as the climate downscaler (ClimGen). A detailed description of the simulation procedures will be provided.

Chapter 4

Impact models and simulation data

The focus of this chapter is to describe the computer models (simulators) and procedure for generating the simulation data. A computer model can be defined as the mathematical equations used to represent the behaviour of the physical system that is being modelled. For instance, in the climate system modelling where a global climate model (GCM) is run to represent key features of the climate system; (atmosphere, ocean and sea ice) for the projection of future climate change.

We built the emulators using simulation outputs from the following models, Model for the Assessment of Greenhouse Gas Induced Climate Change (MAGICC), spatial Climate Generator (ClimGen) and Lund-Potsdam-Jena managed Land dynamic global vegetation (LPJmL).

4.1 MAGICC6 model

MAGICC is a simple carbon cycle climate model that simulates greenhouse gas (GHG) cycles, radiative forcing, and ice melt. The gas cycle uses standard formulae to convert surface emissions of gases to atmospheric concentrations and these, in turn, are then converted to radiative forcing. The generated radiative forcing

is then used to drive a diffusive energy balancing model to estimate global climate change. MAGICC6 is a new version of MAGICC (Meinshausen et al., 2011b) and is able to simulate global mean temperature (GMT) trajectories based on emulation of the seven Atmospheric Ocean General Circulation Models (AOGCMs) used in Solomon et al. (2007) for the Fourth Intergovernmental Panel on Climate Change (IPCC) assessment report.

4.2 ClimGen model

ClimGen is a spatial climate scenario generator. It is based on the pattern-scaling method (which will be described in section 6.4.3) and produces spatial climate change information for a given global-mean temperature change. The method uses the assumption that the pattern of climate change simulated by the coupled AOGCMs is relatively constant but the amplitude changes. These normalised patterns of climate change usually show considerable variation between different AOGCMs, and it is this variation that ClimGen is mainly designed to explore (Osborn, 2009).

The GCM climate change data derived from the pattern scaling method at a 5° by 5° spatial resolution is combined with observations of climate at 0.5° by 0.5° resolution, thereby enabling future climate to be modelled at 0.5° by 0.5° spatial resolution. It facilitates direct coupling between the MAGICC6 and LPJmL models by downscaling the original MAGICC6 results to 0.5° by 0.5° spatial resolution in order to capture detailed representation of spatio-temporal processes involved in the LPJmL model (Osborn, 2009; Mitchell et al., 2004).

One of the downscaling techniques in ClimGen is to incorporate the observed monthly mean climate, and the observed fluctuations in the monthly mean climate (observed time series of deviations), in the future scaled pattern of climate change. The aim is to produce a realistic climate model. The method is used for temperature and precipitation and produces an annual time series that includes natural variability. Precipitation is further modified for inter-annual observed

variability prior to combining it with the climate change pattern. However, wet-day frequency is not calculated directly and instead it is computed from the downscaled precipitation data.

4.3 LPJmL model

The Lund-Potsdam-Jena managed land (LPJmL) model of Bondeau et al. (2007) is a dynamic global vegetation model (Sitch et al., 2003). It uses eco-physiological relations and plant trait parameters for the estimation of photosynthesis, plant growth, maintenance and regeneration loss, fire disturbance, soil moisture, runoff, evapo-transpiration, irrigation and vegetation structure. It generates global vegetation dynamics and the associated carbon and water fluxes.

Agricultural landuse productivity is simulated through varieties of crop functional types (CFTs), both rainfed or irrigated crops. As input variables, the LPJmL model takes climate variables such as precipitation, temperature and insolation. The monthly input and output data are spatially explicit time series of about 60,000 global 0.5° grid cells. Each grid cell can contain a variety of natural or agricultural vegetation types, whose daily growth and productivity is simulated. It derives process-based, large-scale representations of terrestrial vegetation dynamics in terms of plant and crop functional types. A simulation of 100 years takes about 8 hours when the model is run in parallel on 40 processors. But to achieve equilibrium in the carbon pools it is necessary to run the model for about 1000 years prior to the transient time. This is very time consuming (Sitch et al., 2003; Gerten et al., 2004).

In addition, global crop models are tuned to approximate current management practices. In LPJmL, this tuning is based on the maximum Leaf Area index (LAImax) value that can be reached within a growing period and two associated parameters. These parameters encompass the effects of vegetation density, fertilizer application, pests and other factors. Fader et al. (2010) described the implementation of LPJmL in detail and Bondeau et al. (2007) described further

the performance of LPJmL in simulations of crop yields, crop phenology and carbon fluxes.

In summary, MAGICC emulating the global-mean temperature response to forcings. MAGICC6 generates and provides future trajectories of global mean temperature to ClimGen in order to ensure consistency between the global and regional climate simulations. ClimGen emulates the spatial response patterns by disaggregating the temperature trajectories to 0.5° resolution of spatial climate change patterns of air temperature, precipitation, wet day frequency and cloud cover, adding natural variability (weather).

ClimGen generated the climate scenarios which are then supplied as inputs to run the LPJmL simulation for the assessment of climate impacts on variables such as potential crop yields and carbon fluxes. There must be consistency between the spatial scale of LPJmL and climate models and since LPJmL requires climate inputs on a 0.5° by 0.5° grid, the outputs from the global climate models were downscaled using the ClimGen to this spatial resolution. We used CO_2 trajectories (annual concentrations) associated with MAGICCs global temperature trajectories as an additional input to LPJmL.

4.3.1 LPJmL Simulation

LPJmL was run on seven climate change patterns, namely, Canadian Centre for Climate Modelling and Analysis Coupled Global Climate Model (CCCMA-CGCM31), Center for Climate System Research. Two versions of the Model for Interdisciplinary Research on Climate (CCSR-MIROC32HI), CCSR-MIROC32MED and Hadley Centre Global Environmental Model, Met Office United Kingdom (UKMO-HADGEM1), Goddard Institute for Space Studies (GISS-MODELEH), GISS-MODELER, and Institut Pierre Simon Laplace (IPSL-CM4) that had been generated using ClimGen, which used trajectories of global mean temperature constructed by MAGICC. In the MAGICC model, the forcing pathways of all four Representative Concentration Path-ways (RCPs) were used which cover a

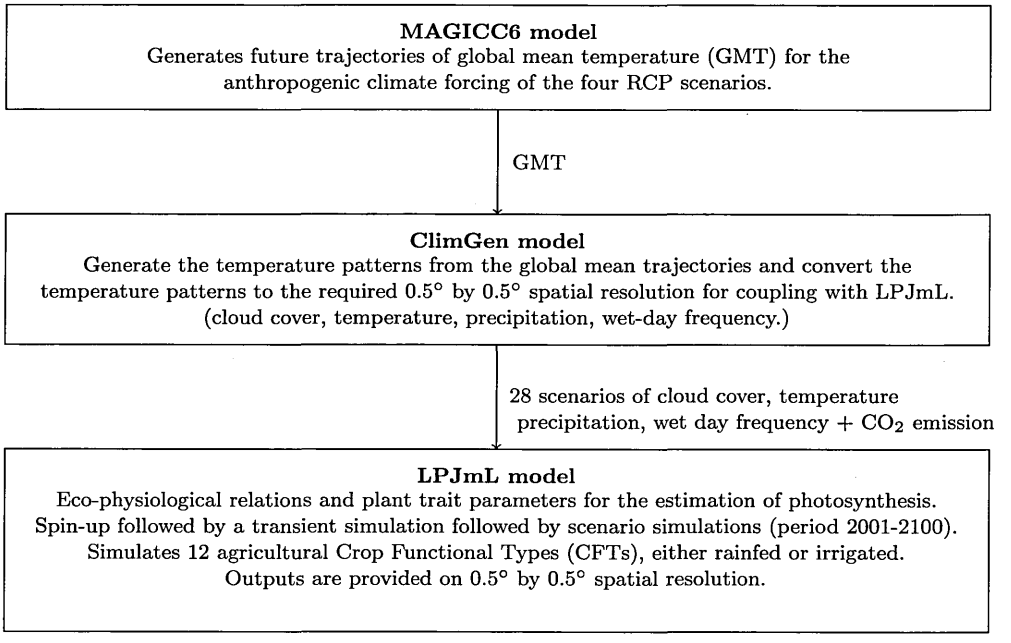


Figure 4.1: Schematic diagram of coupling between climate model (MAGICC6), spatial scenario generator (ClimGen) and impact model (LPJmL) which show key stages for the simulation set up.

range of climate model-structural uncertainty.

The simulations involve a spin-up stage that was used to equilibrate the long-term carbon stores (in natural and agricultural ecosystems) by repeating the observed climate (1901 – 1930 period) 33 times, immediately followed by an additional 13 repetitions in which landuse change were incorporated for the reconstruction of historical soil carbon pools (Fader et al., 2010). Then simulations followed for the period 1931 – 2000, with transient climate and land use change data. The scenario period covers (2001 – 2100) for different RCPs and GCM patterns. Land use change pattern and irrigation were held constant at their year 2000 values. The experiment simulated changes in the carbon fluxes and potential crop yields of each of 12 rainfed and irrigated crop types.

4.3.2 Simulation data

The input data are monthly climate variables (temperature, precipitation, cloud cover and wetday frequency) from ClimGen. All the simulation data are on a global $0.5° \times 0.5°$ degrees resolution. The data cover the period 2001 – 2100.

There are seven different General Circulation Models (GCMs) and each is applied to four RCPs, giving a total of 28 different climate scenarios. The other inputs are annual CO_2 concentrations for all the four RCPs, and soil data that describes 8 different soil types.

The output data from LPJmL are carbon fluxes, namely NPP, FC, HR and potential crop yields. The carbon fluxes are monthly data while the crop yields are annual values for each of 12 CFTs, with rainfed and irrigated outputs considered separately. We have chosen the following CFTs (temperate cereal, rice, maize and oil (soybeans, rapeseed, groundnut, sunflower) for emulation.

In addition, the crop yields have 7 seven different crop management levels for each scenario, and simulations were performed both with and without the CO_2 fertilization effect. In calibrating crop management, the Leaf Area Index (LAI) is a key parameter. LAI is the ratio of total upper leaf surface of vegetation divided by the surface area of the land on which it grows. Crop management levels are represented by maximum leaf area index LAI_{max}. In LPJmL, it is defined as aggregated natural and artificial nutrient availability in combination with other management factors. It represents a proxy for vegetation density (thus reflecting the vegetation response to the overall management intensity). Together with other synchronously varied parameters, it is used to calibrate the modelled yields with respect to observed yields that vary with local management practice, as in Fader et al. (2010). Here we use seven simulations with fixed parameters for each grid cell and crop type so as to derive the yield levels that would be achieved if those management levels were in place.

Different LAI_{max} levels represent different management practices in the LPJmL model. Developed countries are assigned higher LAI_{max} value and developing countries take relatively low values. A low LAI_{max} can be interpreted as a low management practice. The LAI_{max} parameter takes values between 1 and 7 for each CFT and varies with country. A LAI_{max} of 7 corresponds to very high management intensity while LAI_{max} of 1 means a poorly managed system. These

parameters are the same for both rainfed and irrigated simulation data (Fader et al., 2010).

Chapter 5

Methodology

In this Chapter we describe the standard statistical methods used in this thesis. Figure 5.1 gives an overview of the emulation.

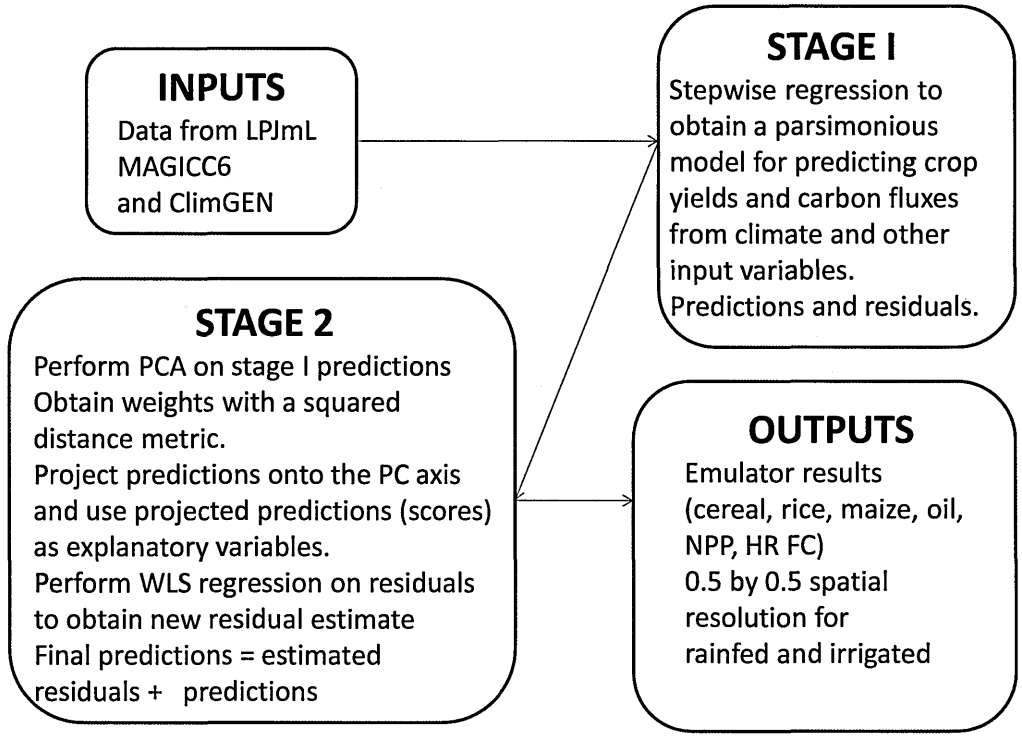


Figure 5.1: Key stages for emulator construction.

5.1 Ordinary Least Squares (OLS) method

A standard statistical task is to predict the value that a dependent variable will take when the independent variable takes some specified value, say x_0.
A univariate regression model is given by

$$Y = X\beta + \varepsilon \tag{5.1}$$

where, Y is a $(m \times 1)$ matrix of observations, m is the number of observations, X is a $(m \times p)$ matrix of predictor variables, p is the number of predictor variables, ε is a $(m \times 1)$ matrix of error terms that are independent and identically distributed, and β is a $(p \times 1)$ matrix of regression coefficients. The least squares estimate $\hat{\beta}$ of β is obtained by minimizing the residual sum of squares

$$\varepsilon'\varepsilon = (y - X\beta)'(y - X\beta) \tag{5.2}$$

which gives

$$\widehat{\beta}_{ols} = (X'X)^{-1}X'y \tag{5.3}$$

where $\widehat{\beta}$ is the least square estimate of β (Yeniay, 2002; Draper et al., 1980; Faraway, 2006; Brown, 1993).

OLS will be applied directly to our data to fit a linear model to explain much of the variation between the response and input variables. We shall use a bi-directional stepwise regression to perform variable selection by combining both forward selection and backward elimination. This method will enable us to obtain a parsimonious model for predicting carbon fluxes and crop yield from seasonal climate and other explanatory variables. The same procedure will be applied to NPP, HR, FC, and crop yields (cereal, rice, maize and oil). These analyses will be done using the Revolution R Enterprise, which has a mechanism for scaling data to handle big computation.

5.2 Model selection criterion

In our analysis, we shall consider quadratic terms with all interactions included. This will result in a large number of regression terms. When there are many parameters to be estimated in a model, it leads to over-fitting. Therefore, it is important to drop all unimportant predictor variables in order to obtain a parsimonious model. For us to achieve that we use stepwise regression. We describe briefly the model selection criteria used in this thesis. There are many criteria for model selection in statistical models such as cross-validation which is a non-parametric method, to parametric methods such as the likelihood ratio test, Akaike and Bayesian Information Criterion (BIC). There is also a Risk Inflation Criterion (RIC) which uses a penalty of $2\ln(p)$ see details in Foster & George (1994).

In statistical inference, the choice of model selection criterion is a critical step towards model improvement. Akaike Information Criterion (AIC) of Akaike (1973) is a prominent selection criterion based on the principles of information theory and it is a reliable model selection method. It selects a model that minimizes the expected error of new or test data, and uses the assumption that all future and past data sets are drawn from the same underlying process and share the same distributional properties as the training data. The best model is the one with the minimum AIC value among the set of models that have been considered. The AIC value is given as

$$AIC = -2\ln(L) + 2p = n\ln(SSE_p) - n\ln(n) + 2p \tag{5.4}$$

where p is the number of parameters in the model (used as a penalty for too many terms in the models), and L is the maximized value of the likelihood function for the model, SSE_p is the sum of squared error for the model and n is the sample size. There is a compromise between the maximized log likelihood and p, the penalty component that measures model complexity (Bozdogan, 2000). A modification

of the AIC technique using a quasi-likelihood was proposed in Pan (2001) and applied to a generalized estimating equation for correlated response data.

Another popular criterion is the Bayesian Information Criterion (BIC) documented in Schwarz (1978). It is a likelihood criterion that also penalizes by the number of parameters to be estimated. It is used as an asymptotic approximation of the Bayesian posterior probability of a given model and is based on the empirical log-likelihood. It penalizes model terms more stringently than AIC. It is given as

$$BIC = -2\ln(L) + p\ln(n) \tag{5.5}$$

The best model in this case is defined as the one that maximizes the BIC value (Schwarz, 1978; Chen & Gopalakrishnan, 1998). The BIC procedure has been applied in various applications, such as linear regression Foster & George (1993). (Posada & Buckley, 2004) compared the performances of AIC and BIC with the likelihood ratio test. Recently, Volinsky & Raftery (2000) applied a BIC criterion on censored data.

Backward elimination: This procedure involves starting with a model that includes all p input variables. At each stage of the procedure, the algorithm computes the partial F-statistic for each variable, assuming it was the last variable to enter the model. The lowest partial F-statistic is compared with a pre-selected significance value F_o. If it is greater than F_o, that variable will be included otherwise the variable will be removed from the model. A new regression model is fitted with $p-1$ variables, the partial F-statistic is computed and the procedure repeated continuously. The algorithm stops when the least partial value is greater than the F_o. The forward selection procedure is similar to the backward elimination except that the model fitting is started with only the intercept term, then a partial F-statistic is computed and compared at each step to the pre-selected F_o value. The term is added to the model provided its partial F-statistic is greater than F_o. The procedure stops when no more terms are important enough to add to the model.

The stepwise regression procedure that combines both backward and forward selection is a modified version of a forward regression that permits re-examination, at every step, of the variables incorporated in the model in previous steps. A variable that entered at an early stage may become superfluous at a later stage because of its relationship with other variables subsequently added to the model. F-statistics are computed and used for determining candidate variables to include or exclude from the model. The stepwise regression in this thesis employs both AIC and BIC to decide which variables are important. In particular, we use AIC to add important terms at each stage of the stepwise process, and BIC to drop irrelevant terms. This procedure is automated and continuous until the AIC or BIC values become smaller which indicates the convergence of the algorithm. The algorithm then stops.

Details of stepwise regression procedure for modelling carbon fluxes are given under sections 6.3 and 6.4.2 of Chapter 6 and subsection 7.1.1 of Chapter 7 for crop yields.

5.3 Censored regression

As stated in section 3.5, censored regression is used when the data on the response variable are limited or it is difficult to observe the full response variable. A censored observation contains partial information about the random variable under consideration. Censored data are common in medical experiments particularly survival analysis. In a clinical trial, when a patient is lost to follow-up (so that the available data provide a lower limit on survival of that patient), or by withdrawing from a treatment such that they can no longer be observed or studied. It also occurs when a subject fails without completing the study or experiment. Similarly, in econometrics for instance, duration of unemployment or lifetime of firms is often censored (Dufresne et al., 2002; Currin et al., 1991; Gosling et al., 2007; Draper et al., 1980).

Censored data are sometimes referred to as a defect of the sample. A closely

related concept is truncation, in which there is a loss of data on both the response and explanatory variables. Truncation involves a greater form of loss of information than censoring. It occurs when observations are incomplete due to some systematic selection process (Field et al., 1998). There are three common forms of censoring namely left-censoring, right-censoring and double censoring. Left-censoring arises when a data point is below a certain value but it is unknown by how much.

In a medical context, if a patient is diagnosed with an illness at time t_1, and dies at time t_2, then the time between contracting illness and death is at least $t_2 - t_1$, but it could be much longer, depending on the time between contracting the illness and diagnosis. Thus, such situations give rise to left censored data.

In contrast, right-censoring occurs when a subject is lost to follow-up or withdrawn before the end of the experiment. For right-censored data in the form of a lifetime variable, suppose that $Y = Y_1, \ldots, Y_k$, is a vector of responses that are right-censored beyond some threshold t_i. Then we shall observe the random variable

$$Z_i = \min(Y_i, t_i) \tag{5.6}$$

and the censoring indicator is given by

$$\delta_i = \begin{cases} 1, \text{if} \quad Y_i \leq t_i \text{ (uncensored)} \\ 0, \text{if} \quad Y_i > t_i \text{ (censored).} \end{cases}$$

We shall observe the random variable for left-censored data and the censoring indicator given by

$$Z_i = \max(Y_i, t_i) \tag{5.7}$$

and

$$\delta_i = \begin{cases} 1, \text{if} \quad Y_i \geq t_i \text{ (uncensored)} \\ 0, \text{if} \quad Y_i < t_i \text{ (censored)} \end{cases}$$

where t_i is the censoring time, which can be a fixed time or a random variable, and δ_i is a censoring indicator showing the event status. For left censored data, observations with values at or below value t are set to t_i.

On the other hand, double censoring occurs when both right and left censoring are present in a dataset. Suppose, $Y_i = \beta X_i + \varepsilon_i$, with $i = 1, \ldots, p$. A variable Y_i is said to be doubly censored if $(Z_i,\ \delta_i,\ X_i)$ is observed instead of $(Y_i,\ X_i)$, such that $Z = \max(\min(Y_i, U_i), V_i)$ and

$$\delta_i = \begin{cases} 1, \text{if} \quad V_i \leq Z_i = Y_i \leq U_i \text{ (uncensored)}, \\ 2, \text{if} \quad Z_i = U_i < Y_i \text{ (right-censored)}, \\ 3, \text{if} \quad Z_i = V_i \geq Y_i \text{ (left-censored)} \end{cases}$$

with $-\infty \leq V_i \leq U_i \leq \infty$ and ε_i is independent of (X_i, U_i, V_i). The situation may arise where infection time may or may not be known exactly, in this case there are may be left-censored data, and the diagnosis procedure can also be fully observed or not (right-censored).

5.3.1 Why use censored regression to analyse crop yield?

OLS is often applied to investigate the relationship between a response variable and its covariates and usually the responses are completely observed and there are no defects in the sample. However, when the responses are not completely observable due to censoring, censored regression is usually considered. Our actual crop yield data from LPJmL can be viewed as censored data. We model a change in crop yield which is a good example of doubly censored data. The change in yield data are characterized by the presence of both positive and negative real numbers with the majority of the data clustered around zero. These zeros in the crop data have physical meaning in the real world. They represent regions where the crop is not currently grown.

Figure 5.2 is an idealized representation of the relationship between crop yield

and temperature. The relationship is well-modelled by the straight line until the yields drop to zero. The zero yield values are not informative, they show that the predicted yield given by the line should not be positive for low temperatures. How to model this? One good option is to treat the values of zero as left-censored observation that take the value zero or less. Censored non-positive predictions for low temperature are then consistent with the data. The regression model will predict values that are less than zero but the extent to which they are less than zero is informative - a predicted value that is just below zero could become a positive yield if the temperature increases by a small amount while a value far below zero will need a large increase in temperature before it becomes a positive yield.

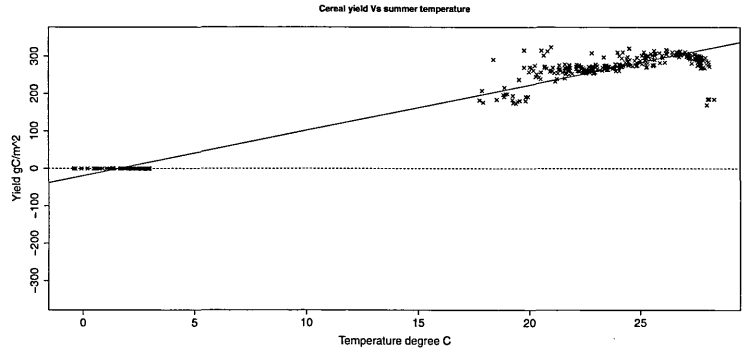
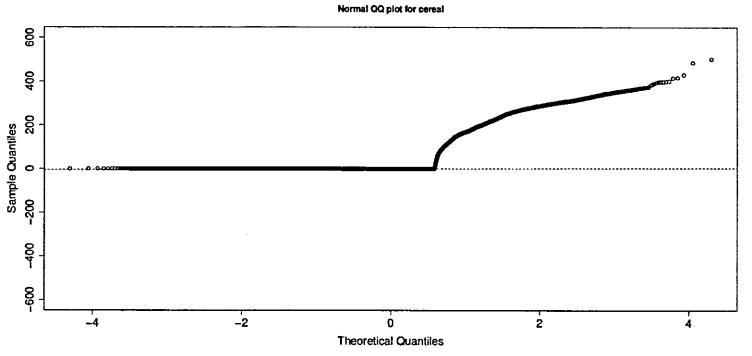


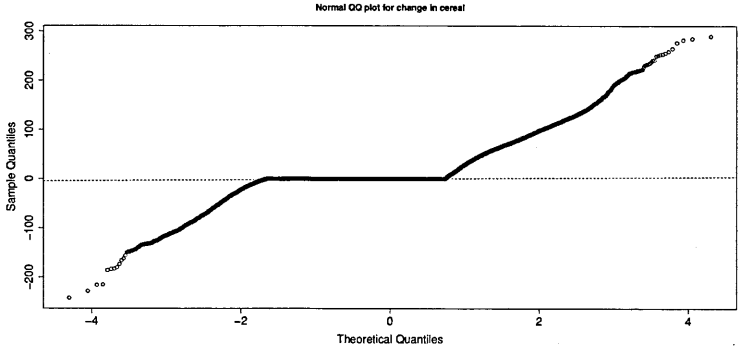
Figure 5.2: Cereal yield Vs summer temperature in (2005-2014) for management level of 5, RCP6 with CO_2 fertilization from CCSR-MIROC32HI.

Figure 5.3 is the normal QQ plot for the actual cereal yield from LPJmL (top) for decade 2005-2014, and for a change in yield (bottom) between decades (2005-2014) and (2085-2094). The positivity of yield gives the data the appearance of being left-censored in the (top) figure. In the lower figure the effect of the zero is more complex. When looking at change in yield between two decades, the yield may be zero in the first decade and/or the second decade, or zero in neither. The consequence is that the change is subject to double censoring.

The censoring indicator is summarized below. Let $Y_{i,t}$ and $Y_{i,t+1}$ each represent $n \times 1$ matrix of crop yield for any two consecutive decades from LPJmL, with $n = 1, \ldots, 59199$, $t = 1, \ldots, 9$ and let $(+)$ sign indicates that crop grows and $(-)$



(a) QQ plot for actual cereal yield in (2005-2014) for management level of 5, RCP6 with CO_2 fertilization from CCSR-MIROC32HI.



(b) QQ plot for change in cereal yield between (2085-2094) and (2005-2014) for management level of 5, RCP6 with CO_2 fertilization from CCSR-MIROC32HI.

Figure 5.3: QQ plots for absolute and change yield.

sign indicates that crop does not grow, then

$$Z_i = \Delta Y = Y_{i,(t+1)} - Y_{i,t} \tag{5.8}$$

$$\delta_i = \begin{cases} 1, & \text{if } Y_{i,t} = + \ \& \ Y_{i,(t+1)} = + \quad \text{(uncensored)} \\ 2, & \text{if } Y_{i,t} = + \ \& \ Y_{i,(t+1)} = - \quad \text{(left-censored)} \\ 3, & \text{if } Y_{i,t} = - \ \& \ Y_{i,(t+1)} = + \quad \text{(right-censored)} \\ 0, & \text{if } Y_{i,t} = - \ \& \ Y_{i,(t+1)} = - \quad \text{(removed)} \end{cases} \tag{5.9}$$

Also, when an observation is censored, we do not know the exact value of the response, thus we do not know the magnitude of error between the response and the regression line relating the dependent to the independent variable. This implies that we cannot use the OLS method to analyse such data (eg crop yield are represented by zeros in areas where the crops are not grown) because we

cannot minimize $\varepsilon'\varepsilon$ to obtain the true parameter estimate.

The censored regression approach captures information on the non-responses (zero observations) in our model. The approach will treat the zeros as censored observations. Censoring is needed because we do not know a priori how unsuitable the climate is for the plant growth. The climatic condition may be close to or far from optimal condition.

It is possible to have the same amount of crop harvest in any two decades (response equal to zero). The zero observed in the change in crop yield for any two successive decades can arise as a result of observing zero in both cases (the crop does not grow in either decade) or when the same amount of crop yield is harvested in a grid cell for two consecutive decades. Censoring is only relevant in the first case and not the latter, in which no threshold-type behaviour has occurred. We use a censored regression, which is a parametric survival model that uses a maximum likelihood for the estimation of its parameters. The procedure is described below.

5.3.2 Maximum Likelihood Estimation (MLE)

A likelihood function is the probability distribution of the observed data but viewed as a function of the model parameters while the data are treated as fixed. It is used to estimate the parameters of the statistical model. Let there be a sample $y_1, y_2, \ldots, y_n$ of n independent and identically distributed observations with a probability density function $f(y|\theta)$ and let the joint probability density function be

$$f(y_1, \ldots, y_n \mid \theta) = f(y_1|\theta) \times \ldots \times f(y_n|\theta) \tag{5.10}$$

Then the log likelihood function is defined as

$$\ln\left(\mathcal{L}(\theta\,;\, y_1, \ldots, y_n)\right) = \ln\left(f(y_1, \ldots, y_n \mid \theta)\right) = \ln\left(\prod_{i=1}^{n} f(y_i|\theta)\right) = \sum_{i=1}^{n} \ln f(y_i|\theta) \tag{5.11}$$

The value of θ that maximizes the likelihood function $L(\theta)$ is denoted as $\widehat{\theta}$. As an example for a Gaussian distribution,

$$f(y_1, \ldots, y_n|\beta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{(2\pi)}} \exp^{-(y_i-\beta)^2/(2\sigma^2)} \tag{5.12}$$

$$= \frac{(2\pi)^{(-n/2)}}{\sigma^n} \exp[-\sum_{i=1}^{n}(y_i - \beta)^2)/(2\sigma^2)] \tag{5.13}$$

$$\ln f = \frac{-1}{2} n \ln(2\pi) - n \ln \sigma - \frac{\sum_{i=1}^{n}(y_i - \beta)^2)}{(2\sigma^2)} \tag{5.14}$$

$$\frac{\partial(\ln f)}{\partial \mu} = \frac{\sum_{i=1}^{n}(y_i - \beta)}{\sigma^2} = 0, \tag{5.15}$$

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{5.16}$$

and

$$\frac{\partial(\ln f)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\sum(y_i - \beta)^2}{\sigma^3} = 0 \tag{5.17}$$

$$\widehat{\sigma} = \sqrt{\frac{\sum(y_i - \widehat{\beta})^2}{n}}. \tag{5.18}$$

The survival function is defined as the probability of an individual surviving beyond time y, and can be given as

$$S(y) = Pr(Y > y) = 1 - F(y) = 1 - Pr(Y \leq y). \tag{5.19}$$

Let $y_1, y_2, \ldots, y_n$ be a random sample of data, and let $f(y)$, $S(C_r)$, $1 - S(C_l)$ denote uncensored, right and left-censored observations respectively. Then the joint likelihood function for double censored data is

$$L = \prod_{i \in U} f(y) \prod_{i \in R} S(C_R) \prod_{i \in L} (1 - S(C_L)) \tag{5.20}$$

where U is the set of uncensored observations, R is set of right-censored data, L is the left-censored data,with $n = U \cup R \cup L$. For right censored data, we observed a pair of random variables (Z, δ), where δ is a censoring indicator represented by

0 (right-censored) and 1 (uncensored) data, and $Z = \arg\min(Y, C_r)$, with t_i in equation 5.6 replaced by C_R here (Lee et al., 1980; Bravo & De Fuentes, 2002; Balakrishnan, 1989). For the case where $\delta = 0$

$$Pr[Z, \delta = 0] = Pr[Z = C_r|\delta = 0] \times Pr[\delta = 0] = Pr(Z > Cr) = S(C_r) \quad (5.21)$$

and when $\delta = 1$, we have

$$Pr[Z, \delta = 1] = Pr[Z = C_r|\delta = 1] \times Pr[\delta = 1] = Pr[Y \leq Z|Y \leq C_r] \quad (5.22)$$

$= [\frac{f(z)}{1-S(C_r)}] \times [1 - S(C_r)]$
which gives by combining the two results

$$Pr(f(z), \delta) = [f(z)]^{\delta}[S(t)]^{1-\delta} \quad (5.23)$$

Therefore the likelihood function for n random samples of pairs (Z_i, δ_i) is given as

$$\mathcal{L} = \prod_{i=1}^{n} Pr(f(z_i), \delta_i) = \prod_{i=1}^{n} [f(z_i)]^{\delta_i}[S(z_i)]^{1-\delta_i} \quad (5.24)$$

where $S(z_i) = 1 - F(z_i)$.

In the case of double censored data, the general likelihood is given as

$$\mathcal{L} \propto [F(z_{r_1+1})]^{r_1} \prod_{i=r_1+1}^{n-r_2} f(z_i)[1 - F(z_{n-r_2})]^{r_2}. \quad (5.25)$$

Suppose $z(\theta) \sim N(\boldsymbol{\beta}, \sigma)$, and let $r_1 = r_2 = 0$, then we shall have a complete sample as in equation 5.11, and estimates of $\boldsymbol{\beta}_0$ and σ are given respectively by equations 5.16 and 5.18. Also, let $r_1 = 0$, then we have right-censored observations where the ML is given in equation 5.24, and $r_2 = 0$, give rise to left-censored observations (Balasubramanian et al., 1992; Balakrishnan, 1996, 2000). Suppose

we fit a parametric censored regression model of the form

$$Z = \mu+\beta_1 x_1+\ldots+\beta_p x_p+\beta_1 x_{1,1}^2+\ldots+\beta_p x_{p,p}^2+\beta_{1,2}x_1x_2+\cdots+\beta_{(p-1),p}x_{(p-1)}x_p+\sigma\varepsilon \tag{5.26}$$

where $\boldsymbol{\beta} = [\mu, \beta_1, \ldots, \beta_p]$ is a vector of regression coefficients and ε is $N(0, \sigma^2)$ and X represents the $n \times p$ matrix of explanatory variables. In any of these cases, $\widehat{\theta} = [\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2]$ can be easily estimated using a numerical algorithm from R statistical package. One major problem with censored regressions is that the likelihoods do not always exist in closed form and as a result, it often takes much time to estimate all the regression parameters using a numerical algorithm. However, we apply censored regression to the crop yield data using *survival* package in R, and some results will be shown in section 7.1 Chapter 7.

5.4 Weighted least squares regression

Weighted least square regression is a generalization of least squares regression. Rather than minimizing a residual sum of squares, as in equation 5.3, instead we minimise the weighted sum of squares

$$\varepsilon'\varepsilon = (y - X\beta)'W(y - X\beta) \tag{5.27}$$

where $\mathbf{W}$ is a diagonal matrix, $\mathbf{W} = diag[w_1, \ldots, w_n]$, and the w_i are non-negative values called weights. The new $\boldsymbol{\beta}_{wls}$ estimate is (Faraway, 2006; Brown, 1993)

$$\hat{\beta}_{wls} = (X'WX)^{-1}X'Wy. \tag{5.28}$$

5.5 Principal component analysis

PCA is a multivariate analysis technique. It is a decomposition of the data matrix $\mathbf{Y}$ into two different matrices $\boldsymbol{\Gamma}$ and $\mathbf{D}$ that captures much of a significant

pattern in the data $\mathbf{Y}$. PCA is charged with determining the variance-covariance structure of the data by using linear combinations of the original variables. It is a statistical principle for reducing a complex dataset to a lower dimension to show simplified structures that underlie it. PCA can be solved in two different ways, namely the singular value decomposition of the data matrix or the eigen decomposition of the covariance matrix.

Consider the matrix $\mathbf{YY}'$ that is required for a PCA, where $\mathbf{Y}$ is a $N \times p$ matrix, where N is the number of observations and p is the number of data variables. Let the covariance matrix $\mathbf{\Sigma}$ be defined as

$$\mathbf{\Sigma} = \frac{1}{n-1}\mathbf{YY}'. \tag{5.29}$$

Then the eigen decomposition theorem gives

$$\mathbf{YY}' = \mathbf{\Gamma\Lambda\Gamma}' \tag{5.30}$$

where $diag(\mathbf{\Lambda}) = \lambda_1, \ldots, \lambda_n$ is a diagonal matrix of associated eigenvalues, $\mathbf{\Gamma}$ is an $N \times N$ orthogonal matrix of eigenvectors and n is the rank of the covariance matrix $\mathbf{\Sigma}$.

Considering the singular value decomposition of matrix $\mathbf{Y}$ that gives

$$\mathbf{Y} = \mathbf{UDV}' \tag{5.31}$$

where $\mathbf{U}$ is an $N \times n$ matrix such that $(\mathbf{U}'\mathbf{U} = \mathbf{I})$ and its columns are orthonormal eigenvectors of $\mathbf{YY}'$. $\mathbf{V}$ is a $p \times n$ matrix $(\mathbf{V}'\mathbf{V} = \mathbf{I})$ of eigenvectors $\mathbf{Y}'\mathbf{Y}$, while $\mathbf{D}$ is a $n \times n$ diagonal matrix which elements are called the singular values of $\mathbf{Y}$ (square root of the eigenvalues of $\mathbf{YY}'$ or $\mathbf{Y}'\mathbf{Y}$).

The columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{YY}'$ while the columns of $\mathbf{V}$ are the eigenvectors of $\mathbf{Y}'\mathbf{Y}$. The diagonal elements of $\mathbf{D}$ are the square root of the eigenvalues of $\mathbf{YY}'$ or $\mathbf{Y}'\mathbf{Y}$. The eigenvalues obtained from the decomposition

are sorted into descending order. The eigenvector with the highest eigenvalue is the most dominant principal component of the data. The principal components (PCs) $\mathbf{P}$ are given by the projection of the data $\mathbf{Y}$ onto the eigenvectors (Abraham & Inouye, 2014; Wold et al., 1987). We use the PCA routine in R to perform the decomposition.

$$\mathbf{P} = \mathbf{YV} = \mathbf{\Gamma D}. \tag{5.32}$$

We use the PCA technique in the second stage of the emulator for interpolation of residual. Residual interpolation estimates residuals using WLS regression. However, there are a large number of observations from the LPJmL emulator. The data matrix for any scenario is a 59199×16 matrix, whose size may make it difficult to observe the true residual patterns. In order to use this large data for residual interpolation, it is necessary to reduce its dimension. Thereby, we form linear combinations of the data matrix using PCA. The procedure of implementing PCA to achieve data reductions in this thesis will be discussed in section 6.4.3 Chapter 6 and section 8.2 of Chapter 8.

5.6 Bayesian regression

Bayesian linear regression is a linear regression using the framework of Bayesian inference. Bayesian computations have extended and broadened the scope of statistical models that can be handled in practice. This is because of the development of Markov Chain Monte Carlo (MCMC). However, if the model errors have a Gaussian distribution and a given form of the prior distribution is assumed, then the posterior distributions of the model's parameters can be obtained analytically, without MCMC method. A major benefit of using Bayesian regression is the provision of a measure of uncertainty in its analysis. Bayesian regression is based upon the theory of the multivariate normal distribution and matrix partitioning.

5.6.1 Multivariate Gaussian distribution

A random variable $\mathbf{z}$ follows a multivariate Gaussian distribution with mean $\mu \subset \mathbf{R^p}$ and covariance matrix $\mathbf{\Sigma}$ if

$$\mathbf{P(z;\mu,\Sigma)} = \frac{\mathbf{1}}{(2\pi)^{\mathbf{p}/2}}|\mathbf{\Sigma}|exp(-\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)), \tag{5.33}$$

where $\mathbf{\Sigma}$ is a symmetric positive definite $p \times p$ matrix. We write $\mathbf{z} \sim N(\mu, \mathbf{\Sigma})$. Also, suppose the random variable $\mathbf{z}$ is partitioned as

$$\mathbf{z} = \begin{pmatrix} z^m \\ z^n \end{pmatrix} \text{ and put } \quad \mu = \begin{pmatrix} \mu^m \\ \mu^n \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma^{mm} & \Sigma^{mn} \\ \Sigma^{nm} & \Sigma^{nn} \end{pmatrix}, \tag{5.34}$$

where the partitioning is conformal. Then $\mathbf{z^m} \sim \mathbf{N}(\mu^\mathbf{m}, \mathbf{\Sigma^{mm}})$ and $\mathbf{z^n} \sim \mathbf{N}(\mu^\mathbf{n}, \mathbf{\Sigma^{nn}})$ respectively. The conditional distributions are

$$\mathbf{z^m}|\mathbf{z^n} \sim \mathbf{N}(\mu_\mathbf{m} + \mathbf{\Sigma^{mn}}(\mathbf{\Sigma^{nn}})^{-1}(\mathbf{z^n} - \mu^\mathbf{n}), \mathbf{\Sigma^{mm}} - \mathbf{\Sigma^{mn}}(\mathbf{\Sigma^{nn}})^{-1}\mathbf{\Sigma^{nm}}) \tag{5.35}$$

5.6.2 Bayesian linear regression

Let a standard linear regression be

$$y_i = \mathbf{x_i}^\mathrm{T}\boldsymbol{\beta} + \varepsilon_i \tag{5.36}$$

where $\boldsymbol{\beta}$ is a $k \times 1$ vector, and the ε_i are independent and identical normally distributed random variables $\varepsilon_i \sim N(0, \sigma_p^2)$. The likelihood for $\boldsymbol{\beta}$, σ^2 given by a vector of observations $\mathbf{y}$ is given by

$$L(\boldsymbol{\beta}, \sigma^2) = \rho(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{p} \rho(y_i|x_i, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{p} \frac{1}{\sqrt{(2\prod)\sigma_p^2}} \exp(\frac{-(y_i - x_i^T\boldsymbol{\beta})^2}{2\sigma_p^2}) \tag{5.37}$$

$$= (2\prod\sigma^2)^{-p/2} \exp\left(-\frac{1}{2{\sigma_p}^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\mathrm{T}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right) = N(\mathbf{X}^T\boldsymbol{\beta}, \sigma_p^2\mathbf{I}) \tag{5.38}$$

maximum likelihood would give unbiased estimates of the regression parameter $\hat{\boldsymbol{\beta}}$, which are also equivalent to OLS estimate in equation 5.3, where $\mathbf{X}$ is the $p \times k$ matrix of input variables and $\mathbf{y}$ is the column $p-$vector $[y_1 \ \cdots \ y_p]^{\mathrm{T}}$. In the Bayesian context, the data are augmented by a prior distribution. This prior information given over the parameters is then combined with the likelihood function using Bayes theorem to give the posterior distribution for the parameters $\boldsymbol{\beta}$ and σ^2.

$$\rho(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) = \frac{\rho(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2)\rho(\boldsymbol{\beta}, \sigma^2)}{\rho(\mathbf{y}|\mathbf{X})} \tag{5.39}$$

Let the prior distribution for $\rho(\boldsymbol{\beta}, \sigma^2)$ be given as $\rho(\boldsymbol{\beta}, \sigma^2) = \rho(\sigma^2)\rho(\boldsymbol{\beta}|\sigma^2)$ which is a conjugate prior by giving the same posterior distribution, such that σ^2 has an inverse-gamma distribution defined as $\rho(\sigma^2) \propto (\sigma^2)^{-(v/2+1)} \exp\left(-\frac{vs^2}{2\sigma^2}\right)$ and $(\boldsymbol{\beta}|\sigma^2)$ has a normal distribution given by $\mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{A}^{-1})$, with $vs^2 = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ and $v = p - k$. The marginal likelihood is then given as

$$\rho(\mathbf{y}|\mathbf{X}) = \int_{d\beta}\int_{d\sigma^2} \rho(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2)\rho(\boldsymbol{\beta}, \sigma^2)d\boldsymbol{\beta}d\sigma^2 \tag{5.40}$$

The posterior distribution is

$$\begin{aligned}\rho(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) &\propto \rho(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2)\rho(\boldsymbol{\beta}|\sigma^2)\rho(\sigma^2) \\ &= (\sigma^2)^{-p/2}\exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] \\ \times(\sigma^2)^{-k/2}\exp&\left[-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{\beta} - \boldsymbol{\mu})\right](\sigma^2)^{-(\alpha+1)}\exp\left(-\frac{b}{\sigma^2}\right)\end{aligned} \tag{5.41}$$

After some algebraic transformation and using 5.38 and 5.40, then we have a product of normal distribution and inverse-gamma distribution:

$$\begin{aligned}\rho(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) \propto (\sigma^2)^{-p/2}\exp&\left[-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\mu}_p)^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{A})(\boldsymbol{\beta} - \boldsymbol{\mu}_p)\right] \\ (\sigma^2)^{-(p+v)/2-1}\exp&\left[-\frac{b + \mathbf{y}^{\mathrm{T}}\mathbf{y} - \boldsymbol{\mu}_p^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \boldsymbol{A})\boldsymbol{\mu}_p + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}}{2\sigma^2}\right]\end{aligned} \tag{5.42}$$

$$= \rho(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X})\rho(\sigma^2|\mathbf{y}, \mathbf{X}) = \mathcal{N}\left(\boldsymbol{\mu}_p, \sigma^2\boldsymbol{A}_p^{-1}\right) \times \mathrm{IG}\left(\alpha_p, b_p\right) \tag{5.43}$$

where $\boldsymbol{A}_p = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{A})$, $\boldsymbol{\mu}_p = (\boldsymbol{A}_p)^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{A}\boldsymbol{\mu})$, $\alpha_p = \alpha + p/2$, $b_p = b + \frac{1}{2}(\mathbf{y}^{\mathrm{T}}\mathbf{y} + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu} - \boldsymbol{\mu}_p^{\mathrm{T}}\boldsymbol{A}_p\boldsymbol{\mu}_p)$, and $b = vs^2/2$.

$$= \rho(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) \propto \rho(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X})\rho(\sigma^2|\mathbf{y}, \mathbf{X}). \tag{5.44}$$

The predictive distribution is then given as

$$\rho(\mathbf{y}|\mathbf{X}) = T_{2\alpha_p}(\mu_p, \frac{1 + \mathbf{A}_p}{\alpha_p})$$

(David, 2002; Murphy, 2007).

Let the prior distribution for $\boldsymbol{\beta}$ be given as $\boldsymbol{\beta} \sim N(0, \Sigma_p^2)$ and for a situation where σ^2 is known. The likelihood function is quadratic in $\boldsymbol{\beta}$, after some transformation by completing the square makes the likelihood becomes normal in $(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$. Let

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}). \tag{5.45}$$

Using 5.38 and 5.40, then we have (the product of two Gaussian distributions is Gaussian distributed)

$$\rho(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) \propto \exp\left[-\frac{1}{2\sigma_p{}^2}(y - X^T\boldsymbol{\beta})^T(y - X^T\boldsymbol{\beta})\right] \times \exp\left[-\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}\right] \tag{5.46}$$

$$\rho(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\frac{1}{\sigma_p^2}\mathbf{X}\mathbf{X}^T + \Sigma_k^{-1})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right] \tag{5.47}$$

The resulting posterior distribution is a Gaussian distribution with mean $\hat{\boldsymbol{\beta}}$ and covariance matrix A^{-1} as $\hat{\boldsymbol{\beta}} = \frac{1}{\sigma_p^2}(\frac{1}{\sigma_p^2}\mathbf{X}\mathbf{X}^T + \Sigma_p^{-1})^{-1}\mathbf{X}\mathbf{y}$, $A = \frac{1}{\sigma_p^2}\mathbf{X}\mathbf{X}^T + \Sigma_p^{-1}$

$$\rho(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}) \propto N(\frac{1}{\sigma_p^2}A^{-1}\mathbf{X}\mathbf{y}, A^{-1}) \tag{5.48}$$

The posterior predictive distribution over the test output $\mathbf{y}^* = f(\mathbf{x}^*)$ for a test input $\mathbf{x}^*$ using 5.48 is equivalent to averaging over all the Gaussian posterior such that

$$\rho(\mathbf{f}^*|\mathbf{x}^*,\mathbf{X},\mathbf{y}) = \int_{d\boldsymbol{\beta}} \rho(\mathbf{f}^*|\mathbf{x}^*,\boldsymbol{\beta})\rho(\boldsymbol{\beta}|\mathbf{X},\mathbf{y})\mathbf{d}\boldsymbol{\beta} = N(\frac{1}{\sigma_p^2}x^{*T}A^{-1}\mathbf{X}\mathbf{y}, \mathbf{x}^{*T}A^{-1}\mathbf{x}^*). \tag{5.49}$$

See further details in (Rasmussen & Williams, 2006; David, 2002; Murphy, 2007).

5.6.3 Gaussian process (GP)

A Gaussian process (GP) is a collection of an infinite number of random variables that has a joint multivariate Gaussian distribution. GP regression under a Bayesian framework assumes a Gaussian prior on functions and the smoothness of the prior depends on the covariance function (Quinonero-Candela & Rasmussen, 2005). GP is a generalization of a Gaussian probability distribution and it is usually specified by mean $m(x)$ and covariance functions $K(x, x')$. GP can capture non-linearity in the interactions between input and dependent variables, thus making it more applicable and flexible than standard linear regression. However, it can be computationally expensive and numerically unstable for large datasets.

GP emulation is based on the Bayesian technique and the experimental design of computer experiments. Its purpose is to predict model outputs at specified input points (Sacks et al., 1989; Santner et al., 2003). A GP emulator gives a probabilistic prediction of the set of outputs that the simulator would produce if it were run for a particular input design. The probabilistic predictions could either be a fully Bayesian posterior distribution where the prediction is a complete probability distribution of the output for some design points, or the Bayes linear approach may be followed in which the emulator provides summary information like expectation (means), variances and covariances of the simulator

outputs (Goldstein & Rougier, 2006; Goldstein, 1995).

A GP emulator assumes that a simulator output is an unknown function $g(.)$. We can then choose a prior distribution for $g(.)$ using the Bayesian approach and update this distribution, with some data obtained from the simulator runs. In particular, $g(.)$ is often assigned to be a Gaussian Process (GP), that is for any set of input $\mathbf{x} = (x_1, \ldots, x_n)^T$ the joint distribution of simulator outputs $\mathbf{g} = (g(x_1), \ldots, g(x_n))$ is a multivariate normal with mean $m(x)$ and covariance functions $K(x, x')$ (Rasmussen & Williams, 2006). Let

$$\mathbf{z} = g(\mathbf{x}) = \mathbf{h}(\mathbf{x})^T\boldsymbol{\beta} + \delta(\mathbf{x}) \tag{5.50}$$

where $\mathbf{h}(\mathbf{x}) = [1, x, x^2, \ldots]$ a vector of known functions that is chosen to reflect the functional form of the simulator output, and $\boldsymbol{\beta}$ is an unknown hyperparameter to be estimated, $\delta(\mathbf{x})$ is a stochastic process with mean zero and covariance function σ^2. The covariance function must be positive definite that is $\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(x_i, x_j) > 0$

$$\mathbf{Cov}[\mathbf{g}(\mathbf{x}), \mathbf{g}(\mathbf{x}')] = \sigma^2\mathbf{C}(\mathbf{x}, \mathbf{x}') \tag{5.51}$$

where $\mathbf{C}(\mathbf{x}, \mathbf{x}')$ is a correlation function with an hyperparameter σ^2 called process variance. Various correlation functions have been used, depending on the assumptions about the function being modelled. We used the correlation function given below

$$\mathbf{C}(\mathbf{x}, \mathbf{x}') = \prod_{j=1}^{n} \exp^{-\alpha_j|\mathbf{x}_j - \mathbf{x}'_j|^\phi} \tag{5.52}$$

where $\alpha_j > 0$, $\forall j$ with $0 < \phi \leq 2$. We chose ϕ as 2, since the GP becomes infinitely differentiable (i.e very smooth). This correlation function is reliable when Gaussian processes are used to analyse outputs of computer experiments. The vector of smoothness hyperparameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$ measure the rate at which the output response changes with a unit change in input and will be

estimated from the data using MLE technique to be described shortly. A function $g(\mathbf{x})$ can be denoted as a GP,

$$\mathbf{g}(\mathbf{x}) \sim GP(\mathbf{m}(\mathbf{x}), \mathbf{K}(\mathbf{x}, \mathbf{x}')). \tag{5.53}$$

Suppose a GP is given as in equation (5.50). GP regression is based on a Bayesian framework. The general principle in Bayesian analysis is to assign prior distributions to the parameters $\boldsymbol{\beta}$ and σ^2 of the GP, and update these prior distributions with respect to some data $\mathbf{z}$, in order to obtain the posterior estimates using Bayes theorem. A major problem with Bayesian analysis is that the inference on the unknown parameters is often analytically intractable. Analytical solutions are only available for simple models that involve a Gaussian observation model or with the use of suitable conjugate priors. In making inference about hyperparameters, the best approach is to integrate over all uncertain parameters but this approach is not often practicable. Therefore, several numerical techniques have been used to facilitate the Bayesian inference with non-analytical models. The most usual technique is to draw samples from the posterior distribution using a Markov chain Monte Carlo (MCMC) method. The limitation with the MCMC algorithm is that the computational cost with a large parameter space is high.

Let $\mathbf{z} = [(z_i, g(x_i))|i = 1, \ldots, n)]^T$ be the simulation data. Given a prior distribution for hyperparameters,

$$P(\boldsymbol{\beta}, \sigma^2|\mathbf{z}) = \frac{P(\mathbf{z}|\boldsymbol{\beta}, \sigma^2)P(\boldsymbol{\beta}, \sigma^2)}{P(\mathbf{z})} \propto P(\mathbf{z}|\boldsymbol{\beta}, \sigma^2)P(\boldsymbol{\beta}, \sigma^2) \tag{5.54}$$

where $P(\boldsymbol{\beta}, \sigma^2|\mathbf{z})$ is the parameter posterior distribution, $P(\mathbf{z}|\boldsymbol{\beta}, \sigma^2)$ is the likelihood, $P(\boldsymbol{\beta}, \sigma^2)$ the prior distribution for hyperparameters. The marginal likelihood or normalising factor $P(\mathbf{z})$ can also be expressed as

$$P(\mathbf{z}) = \int P(\mathbf{z}|\boldsymbol{\beta}, \sigma^2)P(\boldsymbol{\beta}, \sigma^2)\, d\boldsymbol{\beta}\, d\sigma^2. \tag{5.55}$$

Under a full Bayesian framework, the posterior distribution in equation (5.54) and the marginal distribution in equation (5.55) are often difficult to compute for an arbitrary prior distribution as they may not exist in closed form. Conjugate or non-informative priors can be used to overcome this non-analyticity. A conjugate distribution arises when both the posterior and prior distributions come from the same family. Suppose, we want to make predictions at point x^{new} given some training data $\mathbf{z}$, then the joint distribution of the observed values $\mathbf{z}$ and test point z^{new} under a given prior can be obtained. We already know that data $\mathbf{z}$ is

$$\mathbf{z} \sim \mathbf{N}(\mathbf{H}\boldsymbol{\beta}, \sigma^2\mathbf{C}) \tag{5.56}$$

and $z^{new} = g(x^{new})$ have a joint multivariate Gaussian distribution. Then we have

$$z^{new}|\mathbf{z}, \boldsymbol{\beta}, \sigma^2 \sim \mathbf{N}[\mathbf{m}^{\star}(.), \sigma^2\mathbf{C}^{\star}(.,.)]. \tag{5.57}$$

such that

$$m^{\star}(\mathbf{x}) = h^T(\mathbf{x})\boldsymbol{\beta} + \mathbf{c}\mathbf{C}^{-1}[\mathbf{z} - \mathbf{H}(\mathbf{x})]\boldsymbol{\beta}$$

$$C^{\star}(\mathbf{x}, \mathbf{x}') = c(\mathbf{x}, \mathbf{x}') - \mathbf{c}(\mathbf{x})^{\mathbf{T}}\mathbf{C}^{-1}\mathbf{c}(\mathbf{x}'),$$

where $\mathbf{H}^T = [h(x_1), \ldots, h(x_n)]$ and $\mathbf{c}(\mathbf{x})^T = [c(x, x_1), \ldots, c(x, x_n)]$ is the correlation between the training and test points. The correlation matrix of the input (training) design matrix is given by

$$\mathbf{C} = \begin{pmatrix} 1 & c(x_1, x_2) & \ldots & c(x_1, x_n) \\ c(x_2, x_1) & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ c(x_n, x_1) & \ldots & \ldots & 1 \end{pmatrix} \tag{5.58}$$

In order to make predictions, we integrate over the posterior distribution. The

posterior predictive density is given by

$$P(z^{new}|\mathbf{z}) = \int_{d\boldsymbol{\beta}} \int_{d\sigma^2} P(z^{new}|\boldsymbol{\beta}, \sigma^2) P(\boldsymbol{\beta}, \sigma^2|\mathbf{z})\ d\boldsymbol{\beta}\ d\sigma^2. \tag{5.59}$$

As mentioned above it can be difficult to obtain $P(z^{new}|\mathbf{z})$ analytically. The joint prior distribution $P(\boldsymbol{\beta}, \sigma^2|\mathbf{z})$ in equation (5.59) can be further re-expressed using the probability product identity

$$P(\boldsymbol{\beta}, \sigma^2|\mathbf{z}) = P(\boldsymbol{\beta}|\sigma^2, \mathbf{z}) P(\sigma^2|\mathbf{z}). \tag{5.60}$$

Let the prior distribution on the regression parameter $\boldsymbol{\beta}$ and σ^2 be given by a weak prior such that

$$P(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2} \tag{5.61}$$

Substituting equations (5.56) and (5.61) in the posterior distribution in equation (5.54) and with little algebraic manipulation we can infer that $P(\boldsymbol{\beta}|\sigma^2, \mathbf{z}) \sim N(\hat{\boldsymbol{\beta}}, \sigma^2(HC^{-1}H)^{-1})$ and $\sigma^2|\mathbf{z}$ has an inverse-gamma distribution where $P(\sigma^2|\mathbf{z}) = \int P(\sigma^2, \boldsymbol{\beta}|\mathbf{z}) d\boldsymbol{\beta} \sim (n-p-2)\hat{\sigma}^2 \chi^{-2}_{(n-p)}$. $\boldsymbol{\beta}$ and σ^2 are unknown and need to be estimated (given as generalized least squares estimates in equations 5.66 and 5.67).

Integrating over all hyperparameters values weighted by their posterior. Therefore, equation (5.59) can be expressed as

$$P(z^{new}|\mathbf{z}) = \int_{d\boldsymbol{\beta}} \int_{d\sigma^2} P(z^{new}|\boldsymbol{\beta}, \sigma^2) P(\boldsymbol{\beta}|\sigma^2, \mathbf{z}) P(\sigma^2|\mathbf{z})\ d\boldsymbol{\beta}\ d\sigma^2, \tag{5.62}$$

$$\approx \int_{d\boldsymbol{\beta}} \int_{d\sigma^2} P(z^{new}|\boldsymbol{\beta}, \sigma^2) P(\boldsymbol{\beta}|\sigma^2, \mathbf{z}) (n-p-2)\hat{\sigma}^2 \chi^{-2}_{(n-p)}\ d\boldsymbol{\beta}\ d\sigma^2$$

$$= (n-p-2)\hat{\sigma}^2 \chi^{-2}_{(n-p)} \int_{d\boldsymbol{\beta}} P(z^{new}|\boldsymbol{\beta}, \hat{\sigma}^2) P(\boldsymbol{\beta}|\hat{\sigma}^2, \mathbf{z}) d\boldsymbol{\beta}.$$

The values of the integrand can be evaluated noting that the first term in the integral follows from equation (5.57), and also recorgnise that $P(\boldsymbol{\beta}|\hat{\sigma}^2, \mathbf{z}) \sim$

$N(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2(HC^{-1}H)^{-1})$ (product of two Gaussian distribution). Therefore,

$$P(z^{new}|\mathbf{z}) = (n-p-2)\hat{\sigma}^2\chi^{-2}_{(n-p)} \times N[m^{\bullet}(.), \sigma^2 C^{\bullet}(.,.)]. \tag{5.63}$$

where

$$m^{\bullet}(\mathbf{x}) = h^T(\mathbf{x})\hat{\boldsymbol{\beta}} + \mathbf{c}\mathbf{C}^{-1}[\mathbf{z} - \mathbf{H}(\mathbf{x})\hat{\boldsymbol{\beta}}]$$

$$C^{\bullet}(\mathbf{x}, \mathbf{x}') = C^{\star}(\mathbf{x}, \mathbf{x}') + \Big[(\mathbf{h}(\mathbf{x})^{\mathbf{T}}\mathbf{c}(\mathbf{x})\mathbf{C}^{-1}\mathbf{H})(\mathbf{H}^{\mathbf{T}}\mathbf{C}^{-1}\mathbf{H})^{-1}(\mathbf{h}(\mathbf{x})^{\mathbf{T}}\mathbf{c}(\mathbf{x})\mathbf{C}^{-1}\mathbf{H})^{\mathbf{T}}\Big],$$

$$P(z^{new}|\mathbf{z}) = T_{(n-p)}\Big(m^{\bullet}(\mathbf{x}), \frac{(n-p)C^{\bullet}(\mathbf{x}, \mathbf{x}')}{(n-p-2)}\Big). \tag{5.64}$$

where $T_{(n-p)}$ is a student t distribution with $(n-p)$ degree of freedom (Oakley, 1999; Murphy, 2007; Floresa et al., 2014).

MLE estimation of correlation parameters

There are several techniques for estimating correlation hyperparameters of the GP. Maximum likelihood, restricted maximum likelihood and MCMC methods are the standard ones. Maximum likelihood is computationally feasible and depends on the assumption of a multivariate Gaussian distribution. Estimation of the hyperparameter $\boldsymbol{\alpha}$ using the MLE method is described here. Noting that under GP regression, the prior distribution for the data $\mathbf{z}$ is also a Gaussian distribution.

Let the log-likelihood of the parameters be given as

$$L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\alpha}) = -\frac{1}{2}\Big[(n-p)\ log(\sigma^2) + log(det(\mathbf{C})) + (\mathbf{z} - \mathbf{H}\boldsymbol{\beta})^{\mathbf{T}}\mathbf{C}^{-1}(\mathbf{z} - \mathbf{H}\boldsymbol{\beta})/\sigma^{\mathbf{2}}\Big] \tag{5.65}$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]$ is a vector of correlation lengths and $det(\mathbf{C})$ is the determinant of the correlation matrix $\mathbf{C}$. Given $\boldsymbol{\alpha}$ the estimate of $\boldsymbol{\beta}$ is its generalized least squares estimate:

$$\hat{\boldsymbol{\beta}} = (\mathbf{H}\mathbf{C}^{-1}\mathbf{H})^{T}\mathbf{H}\mathbf{C}^{-1}\mathbf{z} \tag{5.66}$$

and, for σ^2,

$$\widehat{\sigma^2} = \frac{1}{(n-p)}\Big[(\mathbf{z}-\mathbf{H}\hat{\boldsymbol{\beta}})^T\mathbf{C}^{-1}(\mathbf{z}-\mathbf{H}\hat{\boldsymbol{\beta}})\Big] \tag{5.67}$$

In order to compute the MLE of $\boldsymbol{\alpha}$, the estimates of $\hat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ are substituted in equation (5.65) and maximized over $\boldsymbol{\beta}$ and σ^2 to obtain

$$L(\hat{\boldsymbol{\beta}}, \widehat{\sigma}^2, \boldsymbol{\alpha}) = -\frac{1}{2}\Big[n\ log(\widehat{\sigma}^2)(\boldsymbol{\alpha}) + log(det(\mathbf{C}(\boldsymbol{\alpha}))) + \mathrm{n}\Big], \tag{5.68}$$

See Santner et al. (2003); Rasmussen & Williams (2006) for further details. The posterior distribution is obtained as $P(\mathbf{z}|\widehat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\alpha}) \sim N[m^{\bullet}(x), \sigma^2 C^{\bullet}(x, x')]$, with posterior mean function given as

$$m^{\bullet}(\mathbf{x}) = h^T(\mathbf{x})\hat{\boldsymbol{\beta}} + \mathbf{c}(\mathbf{x})\mathbf{C}^{-1}\Big[\mathbf{z} - \mathbf{H}(\mathbf{x})\hat{\boldsymbol{\beta}}\Big] \tag{5.69}$$

and posterior covariance

$$C^{\bullet}(\mathbf{x}, \mathbf{x}') = C^{\star}(\mathbf{x}, \mathbf{x}') + \Big[(\mathbf{h}(\mathbf{x})^{\mathbf{T}}\mathbf{c}(\mathbf{x})\mathbf{C}^{-\mathbf{1}}\mathbf{H})(\mathbf{H}^{\mathbf{T}}\mathbf{C}^{-\mathbf{1}}\mathbf{H})^{-\mathbf{1}}(\mathbf{h}(\mathbf{x})^{\mathbf{T}}\mathbf{c}(\mathbf{x})\mathbf{C}^{-\mathbf{1}}\mathbf{H})^{\mathbf{T}}\Big], \tag{5.70}$$

In Chapter 8 of this thesis, we use GP regression for interpolating the residual in the second stage of the crop yield emulator. Some results from GP and its comparison with the WLS for residual interpolation are also provided in Chapter 8.

5.6.4 Kriging

Kriging is a statistical technique that interpolates across space and allows for a spatial lag relationship for systematic and random components. Kriging can incorporate a wide range of spatial relationships to describe the hidden patterns across locations. It is another method for statistical emulation and it is closely related to Gaussian process emulation. Although we do not use this approach to analyse any data in this thesis, the theory and derivation behind it is similar to

GP regression (Cressie, 1993; Sacks et al., 1989). In addition, prior covariance function is an integral part of kriging modelling technique for measuring degree of spatial proximity. We will use various parametric covariance functions under the WLS regression for incorporating correlation parameters into our distance metric. In order to really understand parametric covariance functions, it is essential to discuss kriging, which describes its usefulness in spatial statistics.

Kriging assumes that data $W(\mathbf{x}) = (W(x_1), \ldots, W(x_N))$ arise from a random field defined on the spatial area of interest, such that the mean

$$E(W(\mathbf{x})) = \mu(\mathbf{x})$$

and $cov(W(\mathbf{x})) = \Sigma$. Kriging uses a linear model for interpolating $W(x_0')$ at an unobserved location x_0', taking

$$\hat{W}(x_0') = \sum_{i=1}^{N} \beta_i W(x_i'),$$

for observed locations $x_1', \ldots, x_N'$, where the coefficients or weights β_i are estimated to minimize the variance of prediction error,

$$E[(\hat{W}(x_0') - W(x_0'))^2] = \sum_{i=1}^{N}\sum_{j=1}^{N} \beta_i\beta_j C(x_i', x_j') + Var(x_0') - 2\sum_{i=1}^{N} \beta_i C(x_i', x_0'),$$

subject to

$$E(\hat{W}(x_0') - W(x_0')) = \sum_{i=1}^{N} \beta_i(x_0')\mu(x_i') - \mu(x_0) = 0,$$

where $\mu(x') = E(W(x'))$ and $C(x_1', x_2') = Cov(W(x_1'), W(x_2'))$ is the covariance function of the random field $W(x')$.

There are several types of kriging including simple, universal, ordinary and Bayesian kriging. They differ in the specification of their mean function $\mu(\mathbf{x})$. For instance, simple kriging is based on the assumption that the mean function is a constant that can take different values for different points, universal kriging

takes the form

$$\mu(x) = \sum_{i=1}^{p} \beta_j f_j(x) = \beta^T f(\mathbf{x}),$$

where $\beta^T = (\beta_1, \ldots, \beta_p)$ are unknown regression parameters, and $f(\mathbf{x})^T = (f_1(x), \ldots, f_p(x))$ are known covariates depending on spatial locations. For any form of kriging the nature of Σ is specified through a covariance function relating $Z(x)$ and $Z(y)$, so that

$$cov(Z(x), Z(y)) = \alpha K_\theta(\mid x - y \mid),$$

where $\alpha > 0$ is a scale parameter, and $\theta^T = (\theta_1, \ldots, \theta_q)$ is a vector of real-valued structural parameters for the covariance function (Handcock & Stein, 1993; De Oliveira et al., 1997). The kriging estimate of $Z(x_0')$ is then given by

$$\hat{Z}_\theta(x_0') = k_\theta^T K_\theta^{-1} Z + b_\theta^T \hat{\beta}(\theta),$$

in which $k_\theta^T = (K_\theta(x_0, x_1), \ldots, K_\theta(x_0, x_N))$, K_θ is the $N \times N$ matrix with $(i,j)^{th}$ element $K_\theta(x_i, x_j)$ which is a function relating the covariance function to the distance between pair of locations. $b_\theta = f(x_0) - F^T K_\theta^{-1} k_\theta$, F is the $N \times p$ matrix $\{f_j(x_i)\}$ and

$$\hat{\beta}_\theta = (F^T K_\theta^{-1} F)^{-1} F^T K_\theta^{-1} Z$$

5.6.5 Parametric covariance functions

We investigate the covariance between each pair of scenarios in Chapter 7 to see how it relates to our proposed distance metric in the second stage and to determine if it will improve the efficiency of the estimated WLS regression. Covariance functions are often used in spatial statistics as a measure of similarity. Several covariance functions will be applied in the WLS fitting for residual interpolation in Chapter 7. The Matern model for the covariance function $\gamma(d)$ between two

points at distance $d = x - x'$ apart is defined as

$$\gamma(d) = \frac{\sigma^2}{2^{\kappa-1}\Gamma(\kappa)}(\phi d)^{\kappa}\boldsymbol{\beta}_{\kappa}(\phi d),$$

for $\kappa, \phi, d > 0$, Γ is the gamma function, and $\boldsymbol{\beta}_{\kappa}$ is the modified Bessel function of the second kind of order κ (κ is a smoothness parametèr). The Gaussian covariance model is given by

$$\gamma(d) = \sigma^2 \exp(-\phi d^2) = \sigma^2 \exp(-3\frac{d^2}{\phi^2}),$$

the exponential covariance model is given by

$$\gamma(d) = \sigma^2 \exp(-\phi d) = \sigma^2 \exp(-3\frac{d}{\phi}),$$

and the spherical covariance model is

$$\gamma(d) = \sigma^2(1 - \frac{3d}{2\phi} + \frac{1}{2}(\frac{d}{\phi})^3), d \leq \phi, \text{and 0 otherwise}$$

(Schabenberger & Gotway, 2005; Chiles & Delfiner, 1999; Cressie, 1993; Handcock & Stein, 1993).

5.6.6 Variogram

Distance computations are crucial and often applied in variogram analysis to evaluate the magnitude of a spatial relationship, which might help in specifying priors parameters in Bayesian modelling. Just like a covariance function, a variogram is a reliable tool for determining the degree of spatial dependence for any spatial random process $Z(x)$. It is used to estimate covariance parameters in geostatistics. It can also be used to compare empirical and theoretical models. It is the variance of the difference between field values at two different locations. For a set of random processes Z_i, $i = 1, \ldots, n$ at spatial points $x_1, \ldots, x_n$, the

empirical variogram in this context is defined below as

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{N(h)} |Z(x) - Z(x')|^2 \tag{5.71}$$

where $N(h)$ is the number of pairs of observations at lag h, such that $h = |x - x'|$ is their distance apart. A valid variogram must be symmetric and positive definite as we described under the GP. It is closely related to the covariance function. For a stationary process the relationship is $2\gamma(x, x') = C(x, x) + C(x', x') - 2C(x, x')$, $C(x, x) = C(x', x') = \sigma^2$ (Ecker & Gelfand, 1997; Cressie, 1993).

5.7 Model performance

We assessed the performance of the emulator using other climate models that had not been used to construct the emulator, taking the proportion of variance (ρ) that the emulator explained as a measure of emulator efficiency. For the crop emulators, this proportion was calculated separately for each combination of climate model/irrigation regime/crop. For one combination, let $\bar{y}$ denote the average value given by LPJmL and let $\tilde{\mathbf{y}}^{\star}_{ijktn}$ be the emulator final predictions for the i^{th} RCP, CO_2 fertilization level j, management level k, time slice t and grid point n. Also, let y_{ijktn} denote the actual LPJmL value to which the latter corresponds.

We compute the squared differences between the actual LPJmL values and $\bar{y}$ and also compute the squared differences between the LPJmL values and the emulator predictions. The proportion of the variance in the LPJmL values that is explained by the emulator is

$$\rho = 1 - \left[\frac{\sum_{i=1}^{4} \sum_{j=1}^{2} \sum_{k=1}^{7} \sum_{t=1}^{8} \sum_{n=1}^{N} (y_{ijktn} - \tilde{\mathbf{y}}^{\star}_{ijktn})^2}{\sum_{i=1}^{4} \sum_{j=1}^{2} \sum_{k=1}^{7} \sum_{t=1}^{8} \sum_{n=1}^{N} (y_{ijktn} - \bar{y})^2} \right] \tag{5.72}$$

and the overall cross-validation root mean squared error (RMSE_{CV}) is

$$\text{RMSE}_{CV} = \left(\sum_{i=1}^{4} \sum_{j=1}^{2} \sum_{k=1}^{7} \sum_{t=1}^{8} \sum_{n=1}^{N} \frac{(y_{ijktn} - \tilde{\mathbf{y}}^{\star}_{ijktn})^2}{(4 \times 2 \times 7 \times 8 \times N)} \right)^{1/2}. \tag{5.73}$$

The data are divided into two different groups, namely test and training data for cross-validation of crop yield emulators in Chapters 7 and 8. The training data consist of the two GCMs we used to build the emulators while another five GCMs are used as the test data for validation. This similar procedure is also applied to carbon flux emulators where we left-out some observations from the training GCM data. We used methods described in this chapter to fit our data. Some results for the percentage of variance explained are provided in Chapter 6 for the carbon flux emulators and Chapters 7 and 8 for the crop emulators.

5.8 Sensitivity analysis

In this section, we discuss sensitivity analysis and various forms of sensitivity methods and their derivations. Sensitivity analysis aims to identify the relative importance of variables in a model, that is, the contribution of each variable to the total variance. The main objective of sensitivity analysis is to examine the sensitivity of the output response of a model with respect to the inputs of the model. Sensitivity analysis is a useful tool for understanding the behaviour of a predictive model. For instance, it can help identify variables or parameters that are irrelevant, and such variables can be removed from the final model. It can also assist in a calibration study for investigating the tuning importance of each parameter. Sensitivity analysis can be related to error propagation which is described as the influence that input variability will have on the model outputs. It is also closely related to uncertainty analysis that deals with the quantification of the overall uncertainty associated with the output responses as a result of the model input uncertainties (Monod et al., 2006; Saltelli et al., 2000).

There are two types of sensitivity analysis, namely local sensitivity and global

sensitivity. Local sensitivity focuses on the local impact of the input variables. It is performed by computing partial derivatives of the output variables with respect to the input variables. The problem with a local sensitivity measure is that the input parameters are only allowed to vary within a small interval around their nominal value and this is less reliable when there are many input parameters that can vary simultaneously.

Global sensitivity analysis, on the other hand, distributes output uncertainty to the uncertainty in input variables, based on their probability distributions that cover the entire range of such variables. It involves the whole range of variations and probability distributions of the input variables. A global sensitivity measure can accommodate many input variables simultaneously, unlike a local sensitivity measure. The standard methods for performing global sensitivity are standardized linear regression coefficients (SRC), partial correlation coefficients (PCC) and Morris method. There is also a Monte Carlo Sobol method which computes the indices by decomposing the variance up to a specified order (Saltelli, 2002; Santner et al., 2003). This method computes the first order and total indices with a minimal cost. We shall describe more fully the variance-based sensitivity method of Sobol that we use in this thesis. Variance based methods are model-free because they do not rely on model approximations. They also cover the full ranges of the input factors, but they can be computationally demanding as they require a large number of simulations.

The variance based decomposition technique is based on the Sobol principle (Sobol, 1993). The key principle is to decompose the total output variance D of the function $f(\mathbf{x})$ into summands of increasing dimensionality. The approach is closely related to the analysis of variance (ANOVA). Let the space of the input factor Ω be defined as

$$\Omega^p = (\mathbf{x}|0 \leq x_i \leq 1; i = 1, \ldots, p) \tag{5.74}$$

Then we have

$$f(x_1,\ldots,x_p) = f_0 + \sum_{i=1}^{p} f_i(x_i) + \sum_{1\le i<j\le p} f_{i,j}(x_i,x_j) + \ldots + f_{1,2,\ldots,p}(x_1,\ldots,p). \quad (5.75)$$

It follows from Sobol (1993) that for equation (5.75) to hold, f_0 must be a constant and integrals of every summand over any of its variables must be zero. Therefore, we have

$$\int_0^1 f_{i_1,\ldots,i_q}(x_{i_1},\ldots,x_{i_q})dx_{i_p} = 0 \quad \text{provided} \quad 1 \le p \le q. \quad (5.76)$$

We can infer orthogonality properties from equations (5.75) and (5.76), such that for any $(x_{i_1},\ldots,x_{i_q}) \neq (x_{j_1},\ldots,x_{j_l})$, then

$$\int_{\Omega^p} f_{i_1,\ldots,i_q} f_{j_1,\ldots,j_l} d\mathbf{x} = 0 \quad (5.77)$$

and similarly

$$f_0 = \int_{\Omega^p} f(\mathbf{x})d\mathbf{x} = 0 \quad (5.78)$$

Sobol (1993) evaluated equation (5.75) using multidimensional integrals

$$f_i(x_i) = -f_0 + \int_0^1,\ldots\int_0^1 f(\mathbf{x})d\mathbf{x}_{\sim i},$$

$$f_{ij}(x_i,x_j) = -f_0 - f_i(x_i) - f_j(x_j) + \int_0^1,\ldots\int_0^1 f(\mathbf{x})d\mathbf{x}_{\sim ij}$$

where $d\mathbf{x}_{\sim i}$ is the integration over all variables except x_i and $d\mathbf{x}_{\sim ij}$ integration over all variables except x_i and x_j. Higher order terms can be obtained accordingly. The variance of $f(\mathbf{x})$ is defined as

$$D = \int_{\Omega^p} f^2(\mathbf{x})d\mathbf{x} - f_0^2, \quad (5.79)$$

and partial variance for each term of equation (5.75) is

$$D_{i_1,\ldots,i_q} = \int_0^1 \cdots \int_0^1 f^2_{i_1,\ldots,i_q}(x_{i_1},\ldots,x_{i_q})dx_{i_1},\ldots,x_{i_q} \quad (5.80)$$

$1 \leq i_1 < \ldots < i_q \leq p$, and $q = 1, \ldots, p$. The estimates of f_0, D and D_i are given respectively as $\widehat{f_0} = \frac{1}{n}\sum_{m=1}^{n} f(\mathbf{x}_m)$, $\widehat{D} = \frac{1}{n}\sum_{m=1}^{n} f^2(\mathbf{x}_m) - \hat{f}_0^2$, and

$$\widehat{D}_i = \frac{1}{n}\sum_{m=1}^{n} f(\mathbf{x}^1_{(\sim i)m}, x^1_{im}) f(\mathbf{x}^2_{(\sim i)m}, x^1_{im}) - \hat{f}_0^2,$$

n is the sample size used to generate MC estimates, $\mathbf{x}_m$ is a sample point $\in \Omega^p$, and $\mathbf{x}_{(\sim i)} = (x_{1m}, x_{2m}, \ldots, x_{(i-1)m}, x_{(i+1)m}, \ldots, x_{pm})$.

If we square and integrate equation (5.75) over Ω^p and use the orthogonality property in equation (5.77), then

$$D = \sum_{i=1}^{p} D_i + \sum_{1 \leq 1 < j \leq p} D_{ij} + \ldots + D_{1,2,\ldots,p} \tag{5.81}$$

and sensitivity measures are given as

$$S_{i_1,\ldots,i_q} = \frac{D_{i_1,\ldots,i_q}}{D} \quad 1 \leq i_i < \ldots < i_q \leq p \tag{5.82}$$

where S_i is the first order sensitivity index for variable x_i and measures the main effect, while S_{ij} is the second order index that measures the interactive effect provided $i \neq j$ (Saltelli et al., 2000).

The general steps taken for performing sensitivity analysis involve:

(i) Design of experiment to identify questions to answer.

(ii) Determine the relevant input factors.

(iii) Assign the probability density to each input variable.

(iv) Generate a sample from each input distribution.

(v) Evaluate the model and compute the relative measure of importance of each input variable on the output response.

Suppose our model is represented by $y = f(X)$ as defined in equation 6.1.

The indices are given respectively as

$$S_i = \frac{Var[E(y|x_i)]}{Var(y)} \tag{5.83}$$

and

$$S_{Ti} = \frac{Var(y) - Var[E(y|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots x_p)]}{Var(y)} \tag{5.84}$$

where $Var[E(y|x_i)]$ is the first order effect of variable x_i, and $Var(y)$ is the total variance of y. We followed this procedure for our sensitivity analysis in Chapters 6 and 7.

The general procedure to compute the indices is to choose a probability distribution for the input variables (sampling and resampling matrices). We use this approach for the carbon fluxes. However, for the crop yields, we sample directly from the simulation data since we have a large crop data. We randomly sample 20,000 observations across all the simulation scenarios but only from simulation with CO_2 fertilization effects. This sample data will be used to compute the Monte Carlo estimates of the sensitivity indices. We present some sensitivity results for the carbon and crop yield emulators in Chapter 6 and 7 respectively. These results will provide hints on how variations in input variables or parameters affect the response variables.

Chapter 6

The emulation of carbon fluxes

This chapter will deal step by step with the procedures and techniques that we have adopted to emulate LPJmL output data. We focus on the emulation of carbon fluxes and construct emulators using varying techniques both for low resolution data ($2° \times 2°$ grid cells) and high resolution data ($0.5° \times 0.5°$ grid cells). The procedures are described below. We also present results from each of the techniques. Important questions we would like to answer at the end of this chapter are:

1 Can LPJmL output (vegetation) be predicted for arbitrary CO_2 scenarios?

2 Can LPJmL output be predicted for arbitrary climate input?

6.1 Emulation of carbon fluxes

The main objective in this section is to model the global terrestrial biospheric response to climate change and anthropogenic CO_2 concentration by constructing a statistical emulator. This is necessary to facilitate the coupling between climate and impact models. Here, we assess and demonstrate the feasibility of taking output data from one model and using it as input data for another model. The results from this chapter give an insight into the feasibility of doing this for LPJmL land-use impact model, when its input variables are climate variables

from the ClimGen model.

In this chapter, we adopt a linear model for the global prediction of carbon fluxes from LPJmL simulations. Two different datasets are analysed. In particular, we investigate the spatial relationship between NPP, HR, FC, climate variables and CO_2 concentration for the first dataset. We also consider a regional model by clustering the data into five world regions. We construct an individual emulator for each region (although these results are not shown). We utilize a similar approach to construct emulators for the second dataset that represents a future scenario dataset with four different RCPs, namely RCP 2.6, 4.5, 6.0 and 8.5.

In this chapter, we consider two different approaches for the modelling of the carbon fluxes, namely one-stage using OLS (low resolution data) and a two-stage technique using a combination of OLS, PCA and WLS (high resolution data).

6.2 Simulation data used for this analysis

We use data from the LPJmL simulation output, which include monthly net primary productivity (NPP), heterotrophic respiration (HR) and annual fire carbon (FC) on a $0.5° \times 0.5°$ resolution over land. Section 6.3 deals with the procedure for emulating NPP on low resolution. Both heterotrophic respiratory and fire carbon loss will be analysed in section 6.4 of this chapter as they are important determinants to complete the link from emissions to vegetation.

The climate data are provided by the MAGICC6 and ClimGen models discussed in Chapter 4. We consider two different simulation datasets in this chapter. The first dataset is monthly data from 1901-2000 that describes the past climate. The input variables are surface temperature, precipitation, wet-day frequency, vapour pressure, diurnal temperature range, near surface temperature maximum and near surface temperature minimum. The second dataset involves future scenarios from 2001-2100, corresponding to possible future emission projections. The climate variables in the second dataset are surface temperature,

precipitation, cloud cover, wet-day frequency and anthropogenic CO_2 concentration (which is annual data). There are four different RCP's and four GCMs in the second dataset. There are 59199 grid cells in the data. The two sets of variables are both valid input variables for the emulation of the LPJmL simulation data, even though the inputs to the simulator itself are the same in both cases (temperature, precipitation, cloud cover and wet day frequency). The three additional variables of near surface temperature minimum, near surface temperature maximum and diurnal temperature range (that are not provided to the 21st century emulator) are summary statistics, which can be derived from simulator input variables. Summary statistics are interesting to consider as inputs as they have been used to model the effects of climate change on terrestrial ecosystems in the literature.

In section 6.3, we shall consider only one GCM (UKMO-HADGEM1) while in section 6.4 we shall use all four GCMs. Our analyses involve all the four RCPs. The reason for the inclusion of several RCPs in our model is to incorporate information on different emission profiles. Their inclusion gives a broad range of emission pathways.

6.3 Emulation of NPP using OLS: low resolution

Carbon flux emulators include emulators for NPP, HR and FC. Here, we present the results for an NPP emulator that uses a low resolution so as to reduce the dimension of the data and make the data more manageable for emulation. The dimension of the grid cell was increased to $2° \times 2°$ resolution, which made handling and processing of the data much easier. We chose to handle the large temporal variability that occurs at high spatial resolution by averaging the monthly data to decadal values. The average decadal NPP given by LPJmL was computed for each $2° \times 2°$ grid cell. We then obtained the change in NPP between every two

successive decades. We applied the change in seasonal climate, initial seasonal climate and latitude as explanatory variables in an ordinary least squares regression. We included CO_2 concentration as an additional input to the 2001-2100 data. The input variables are listed in Table D.1 of Appendix D except that the variables LAI and soil were not included. We left-out 500 observations from each decade for cross-validation purposes.

Before we fitted the regression, the descriptive statistics for decadal NPP (Table 6.1) were computed for the first dataset. This enables the distribution of NPP to be examined. Our analysis involves prediction of the global average NPP change for 10 time-slices from 1901-2000 as a function of the input variables. We also predicted mean decadal change in NPP between the first and last decade, as well as regional predictions by clustering the data into five world regions that are each emulated separately.

To construct our emulator for the prediction of NPP change for all decades, we model the LPJmL simulation output by fitting a quadratic model to the data:

$$\mathbf{y} = f(\mathbf{X})+\varepsilon = \beta_0+\beta_1 x_1+\ldots+\beta_p x_p+\beta_1 x_{1,1}^2+\ldots+\beta_p x_{p,p}^2+\beta_{1,2} x_1 x_2+\cdots+\beta_{(p-1),p} x_{(p-1)} x_p+\varepsilon \tag{6.1}$$

where $\mathbf{y}$ is the LPJmL simulated mean decadal NPP change, p is the number of parameters for estimation and $x_1, \ldots, x_p$ are independent variables. They are seasonal climate variables, CO_2, and CO_2 change. We assume $\varepsilon \sim N(0, \sigma^2)$.

We built the emulator using the *MASS* package in R (2013) and combined this with a stepwise algorithm to reduce the number of terms in the model, as in Holden et al. (2010a,b). This is similar to Box (1988) who earlier used a regression-based approach to establishing an empirical relationship between climate and NPP. For the variable selection procedure, both forward and backward stepwise regressions were used to fit the model. In the forward stepwise mode, we fitted a linear model (starting from a null model), and variables were included step by step until the algorithm selected about 250 terms in the model. AIC

was used as the stopping rule criterion. This process was followed by backward elimination using BIC where non-significant variables are removed step by step until BIC would be reduced by any further removal. BIC is a more stringent criteria for variable selection than AIC so the number of terms in a model was reduced substantially.

Having built the emulators, we also performed a sensitivity analysis to examine the relative contribution of each input variable in our respective models. As explained in section 5.8 of Chapter 5, calculating the total effects of each explanatory variable helps identify the relative importance of variables in a model, that is, the contribution of each variable to the total variance. We use the Sobol global sensitivity method. This computes the indices by decomposing the variance up to a specified order. The method we used computes the first order and total indices with a minimal cost. Suppose our model is represented by $y = f(\mathrm{X})$ as defined in equation 6.1. The indices are given respectively by equations (5.83) and (5.84) where $Var[E(y|x_j)]$ is the first order or main effect of variable x_j, and $Var(y)$ is the total variance of the response y. S_{Tj} is the result of the main effect of x_j and all its interaction with other parameters up to order p (Saltelli, 2002; Santner et al., 2003).

We assumed that all our input variables were from a Gaussian distribution; we generate a Monte Carlo sample of 20000 observations to estimate the sensitivity indices. This procedure is automated in the *sensitivity* package of R (2013).

In Table 6.1, the minimum decadal NPP is -1.3920 gCm^{-2} so in 1931-1940, in that decade, the total respiratory carbon loss exceeded the total carbon fixed by the plants for a particular grid cell. The median NPP is the same for the first two decades at 28.52 gCm^{-2}. Similarly, mean decadal NPP is the same for decades 1 and 2, both the NPP decadal mean and median are increasing with time. The last decade in the century has the highest median and mean NPP (31.8 and 33.53 gCm^{-2}) respectively. The maximum NPP for the century occurs in 1961-1970 with 145.8 gCm^{-2}. The data are fairly skewed, we could have transformed the

Table 6.1: Summary of mean decadal NPP in gCm^{-2} for 10 time-slices from 1901-2000. The results show the mean and quartiles

Year	Min	1st Qu.	median	mean	3rd Qu.	max
1901-1910	0.7967	6.6290	28.5200	29.9700	43.6500	123.2000
1911-1920	-0.7967	6.6290	28.5200	29.9700	43.6500	123.2000
1921-1930	-0.7537	7.0390	28.5500	30.1700	43.8200	117.6000
1931-1940	-1.3920	7.1400	29.1300	30.4300	43.7600	130.4000
1941-1950	-0.7262	7.5730	29.3400	30.8400	44.7500	139.9000
1951-1960	-0.7262	7.5730	29.3400	30.8400	44.7500	139.9000
1961-1970	-0.5127	8.2390	29.0500	31.2500	45.2900	145.8000
1971-1980	-0.1701	8.9090	30.3800	32.3500	46.1100	142.9000
1981-1990	-0.6856	9.2290	30.7000	32.6200	46.9200	144.6000
1991-2000	-0.5185	10.8500	31.8000	33.5300	47.8000	138.7000

Table 6.2: Proportion of variance ρ obtained from the prediction of mean NPP changes for successive decades, using seasonal climate change, baseline climate and CO_2 as inputs. The stepwise algorithm selects about 180 terms in the model.

	ρ for	ρ for 2001-2100			
	1901-2000	RCP3	RCP4.5	RCP6	RCP8.5
Decades (1 & 2)	0.74	0.94	0.95	0.94	0.89
Decades (2 & 3)	0.78	0.96	0.96	0.95	0.82
Decades (4 & 5)	0.79	0.94	0.95	0.95	0.91
Decades (9 & 10)	0.73	0.92	0.93	0.93	0.91
All time-slices					
All decades	0.79	0.81	0.83	0.82	0.52

data represent a normal sampling.

Table 6.2 (above) shows the ρ values from various quadratic models fitted to each decadal change in NPP for the two datasets. For the period 1901-2000, these models explained at least 73% of the variance in the response variable, and the best model explained 79% of the response variance for NPP change in decades (9 to 10) and (4 & 5) respectively (Table 6.2). The result of decades 4 & 5 is further examined in Figures 6.1 and 6.2. For the future scenario data with four RCP's, the response variance of NPP under RCP 8.5 is explained less well than for other RCP's. We have ρ as high as 0.96 for the predicted change in NPP between decades (2 & 3) for RCP 3 and RCP 4.5. The emulation results are generally much better for the future simulations than for the past simulations, probably because the changes in carbon flux under future scenarios are larger, increasing the signal to noise ratio relative to the past climate data.

The bottom row of Table 6.2 shows the results for a model fitted to the aggregated data that is, fitting a single model to the entire data. This procedure enables us to include CO_2 and CO_2 change as inputs into the model. The CO_2 can only be included as an input to the model by aggregating all the data because the CO_2 data are not spatially resolved; they are a global average. The results are generally good with $\rho \geq 0.80$ except for RCP 8.5 where $\rho = 0.52$. The model fitted to individual decadal data is consistently better than the joint model, as perhaps one would expect.

Figure 6.1(a) shows diagnostic plots for the fitted model used for predicting mean decadal change in NPP between decades 4 & 5. It indicates that our fitted models (emulators) satisfy the basic assumptions of linear regression. First, the fitted vs. residual plots are evenly scattered, showing no evidences of heteroscedality. Second, the normal QQ-plot (top-right) is almost a perfect straight line, suggesting that the assumption of normality is not seriously violated. The scale-location plot of the standardized residual against the fitted values also indicates that most of the points are within 2 standard deviations of the regression

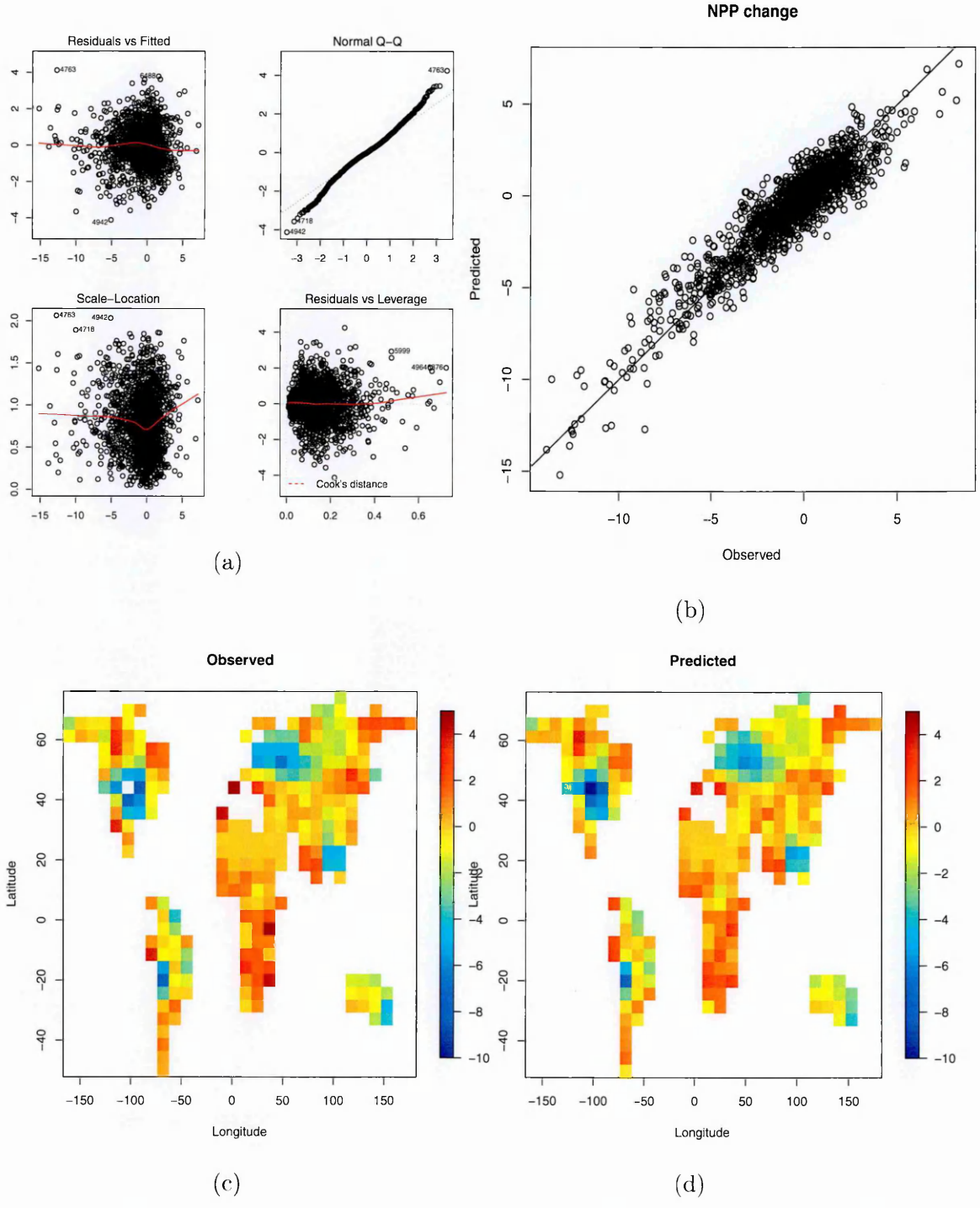


Figure 6.1: Mean decadal NPP change in gCm^{-2} between (1931-1940) & (1941-1950): (a) diagnostic plots; (b) pair plot between observed and predicted change in NPP; (c & d) spatial plots of observed and predicted NPP change. The observed values (c) are the simulated NPP values given by the LPJmL model while the predicted values (d) are the given by the emulator

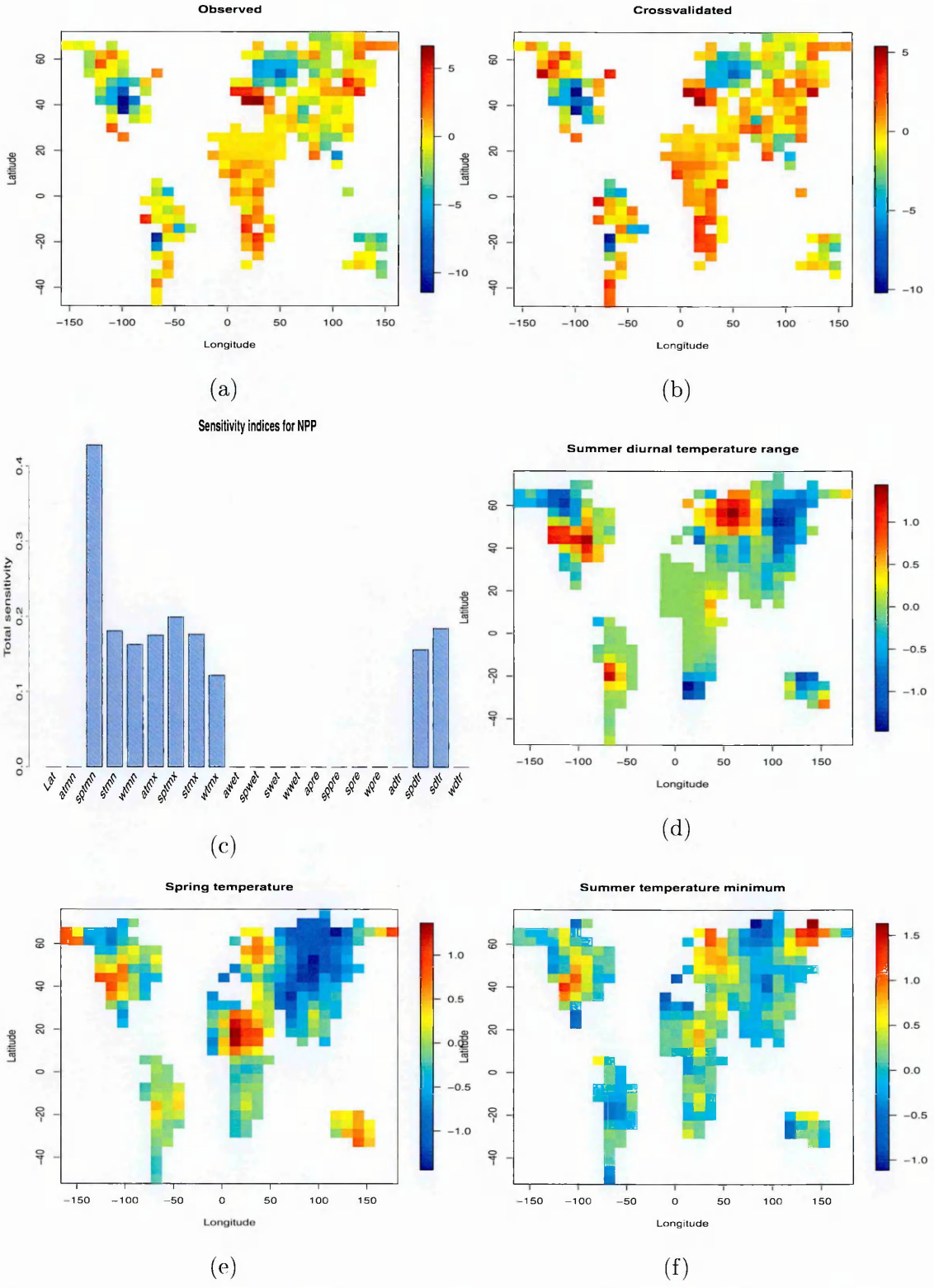
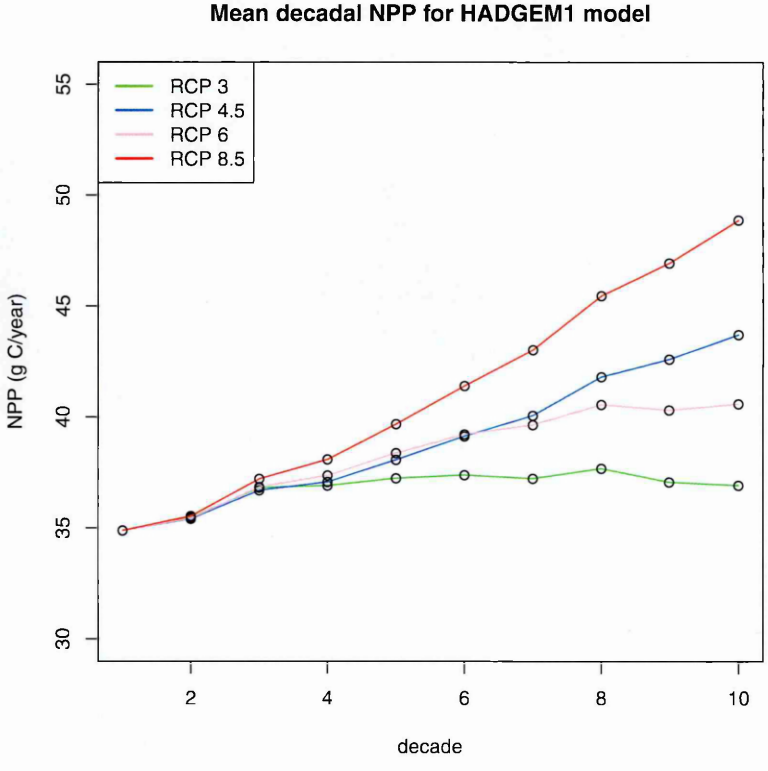


Figure 6.2: Mean decadal NPP change in gCm^{-2} between (1931-1940) & (1941-1950): (a) & (b) spatial maps for the 500 points omitted for cross-validation, observed (a) and validated (b) respectively; (c) sensitivity index for variable importance; (d, e & f) are spatial plots of the 3 most important predictors, namely summer diurnal temperature range (d), spring temperature (e) and summer temperature minimum (f).

Figure 6.3: Plots of the long-term mean decadal NPP for the four RCPs future scenarios (2001-2100) from UKMO-HADGEM1.



line while the bottom-right plot identifies few visible outliers in the large dataset.

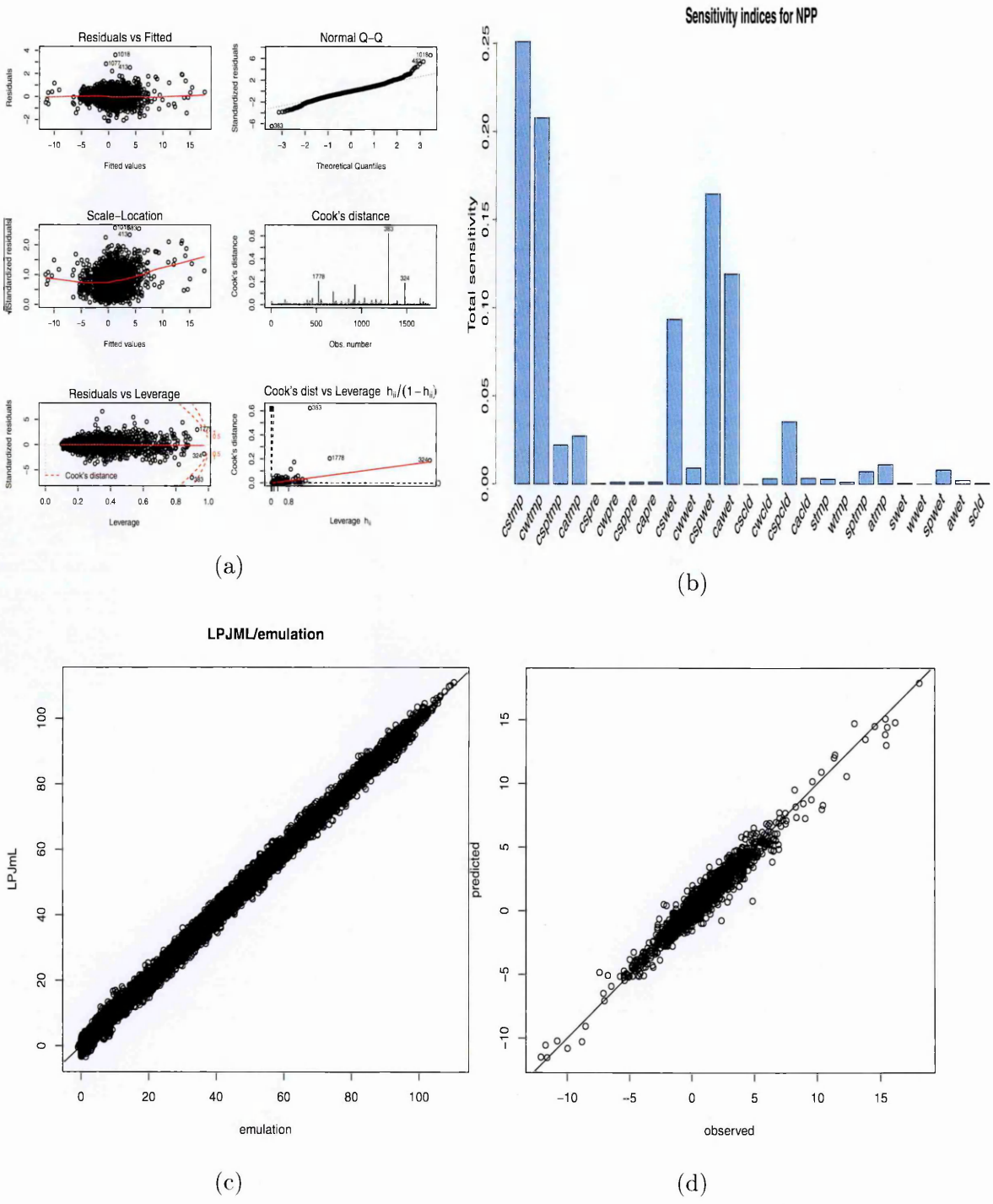
Figure 6.1(b) is a pairwise plot of the observed NPP (LPJmL simulated) and the predictions given by the emulator. The plot clearly shows a high correlation between them; there is a correlation of 0.92 between the observed and predicted values. Figure 6.1(c & d) are the spatial plots of observed and predicted change in NPP. The observed global pattern of NPP is very similar to our model's predictions.

Figure 6.2(a & b) are spatial plots comparing the emulator with LPJmL for the 500 observations left-out for cross-validation. The results indicate that the model does relatively well in most regions. Figure 6.2(c) is a sensitivity index plot for the relative contribution of each predictor in the model. This identifies variables that are more relevant for predicting the true dynamics of change in NPP between decades 4 & 5. The model identifies the three most important predictors as summer diurnal temperature range (sdtr), summer temperature minimum (stmn) and spring temperature (sptmp). Although these significant variables are absent in the future data, they are correlated with temperature which is present in the 21st century data. As mentioned under section 6.2, these three variables can be derived from other input variables.

Their spatial distribution is given in Figures 6.2(d-f). Overall, temperature seems to be the most important factor for determining the global NPP level.

Figure 6.3 shows the long-term mean decadal NPP for the four RCPs future scenarios. It gives a plausible description of the future NPP pattern under future climate change and CO_2 emissions. The plot shows a low NPP level for RCP3. The RCPs 4.5 and 6 scenarios give rise to a slight increase in NPP level. For RCP 3, NPP level is increasing but very slowly while for RCP 8.5 the increase is about an order of magnitude greater.

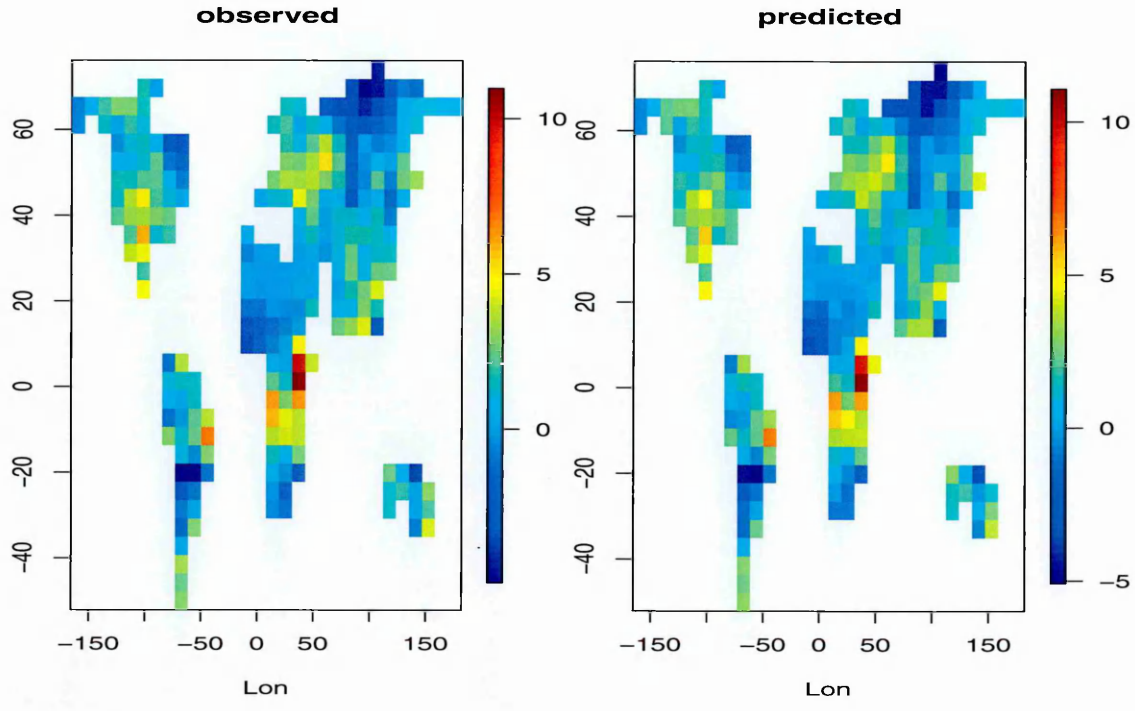
Figure 6.4 shows the results of the model for the prediction of mean decadal change in NPP between decades 2 & 3 for RCP3. Figure 6.4(a) is the diagnostics plot for checking the model assumptions (cf Figure 6.1(a)). None of the assump-



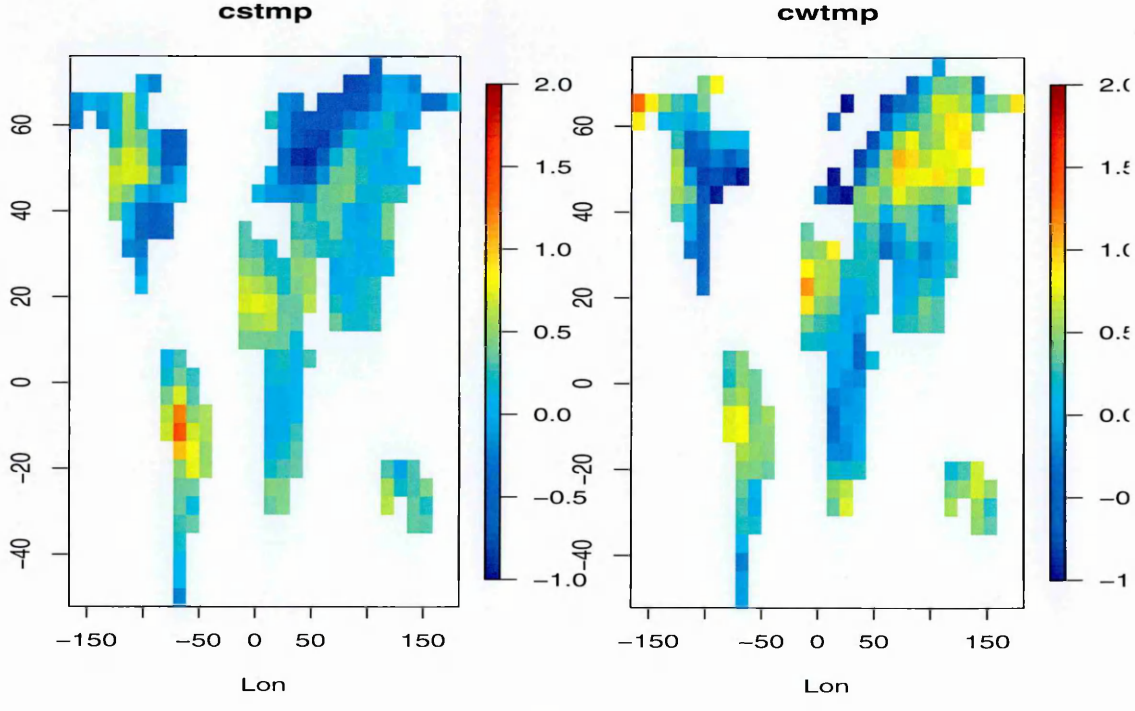
(a) (b)

(c) (d)

Figure 6.4: Diagnostic plots for the mean decadal change predictions of NPP in gCm^{-2} for period between 2011-2020 & 2021-2030 for RCP3; (b) sensitivity index for important variable; (c) pairwise plot between observed (LPJmL) and predicted (emulated) NPP change for the whole century; (d) pairwise plot between observed and predicted NPP change.



(a)



(b)

Figure 6.5: Spatial plots for predictions of mean decadal change in NPP between 2011-2020 & 2021-2030 for RCP3, (a) observed and predicted NPP change (gCm^{-2}); (b) change in summer temperature (cstmp) and change in winter temperature (cwtmp) in °C. The plots of other significant variables are not shown.

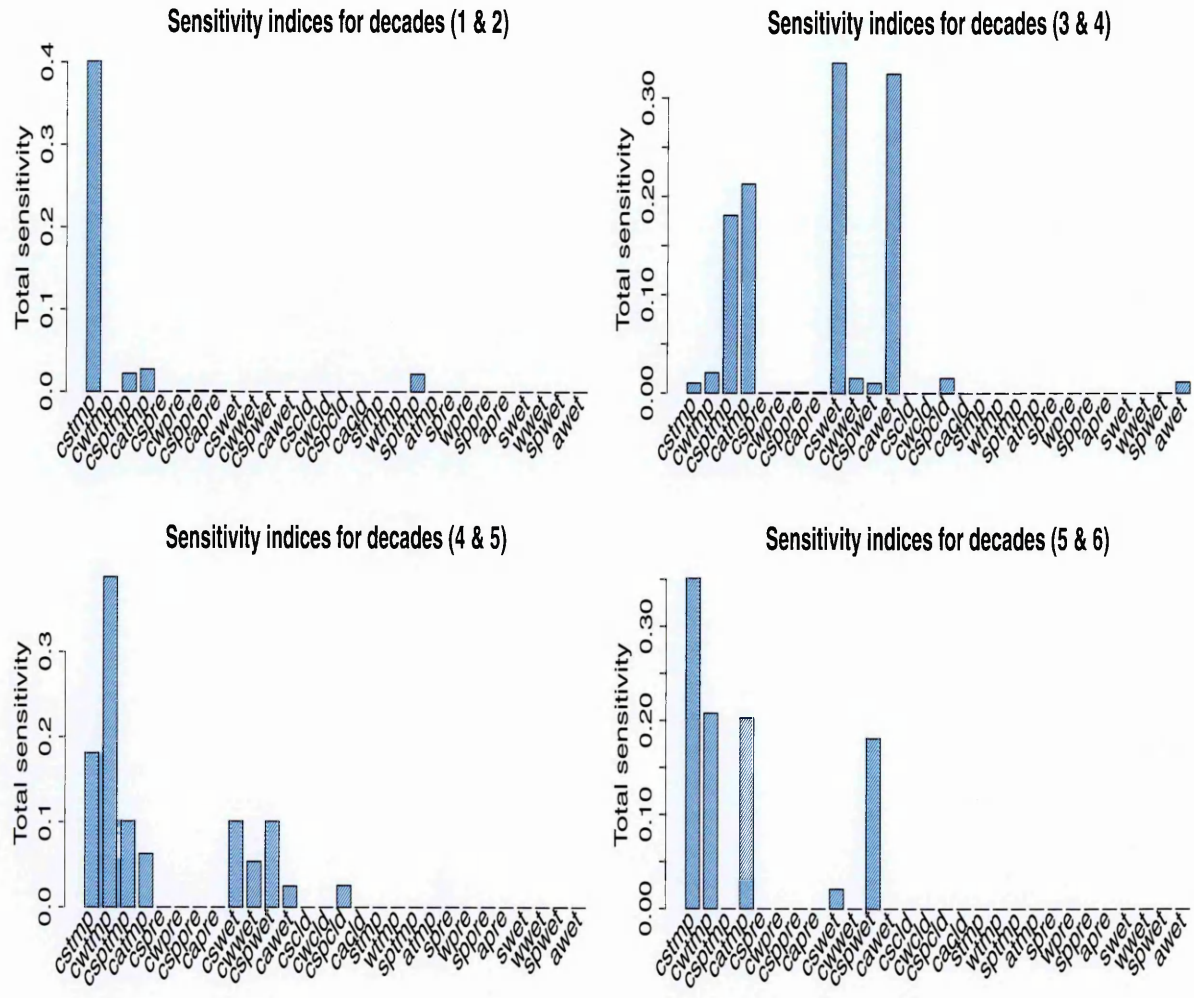


Figure 6.6: Sensitivity index for the predicted mean decadal NPP change between decades (1 & 2) top-left, (3 & 4) top-right, (4 & 5) bottom-left and (5 & 6) bottom-right plots for RCP3. Note: each decade of data is modelled independently.

tions seem to have been seriously violated by our fitted model. The residual vs. fitted plots are uniformly scattered, and the QQ plot is almost a straight line, except for the end of the plot.

Figure 6.4(b) is the sensitivity index for the fitted model and identifies the four most important predictors. These were change in summer temperature, change in winter temperature, change in spring wetday frequency and change in autumn wetday frequency. Thus, in this model, temperature and wet-day frequency are the important climate variables for the prediction of global mean decadal NPP change.

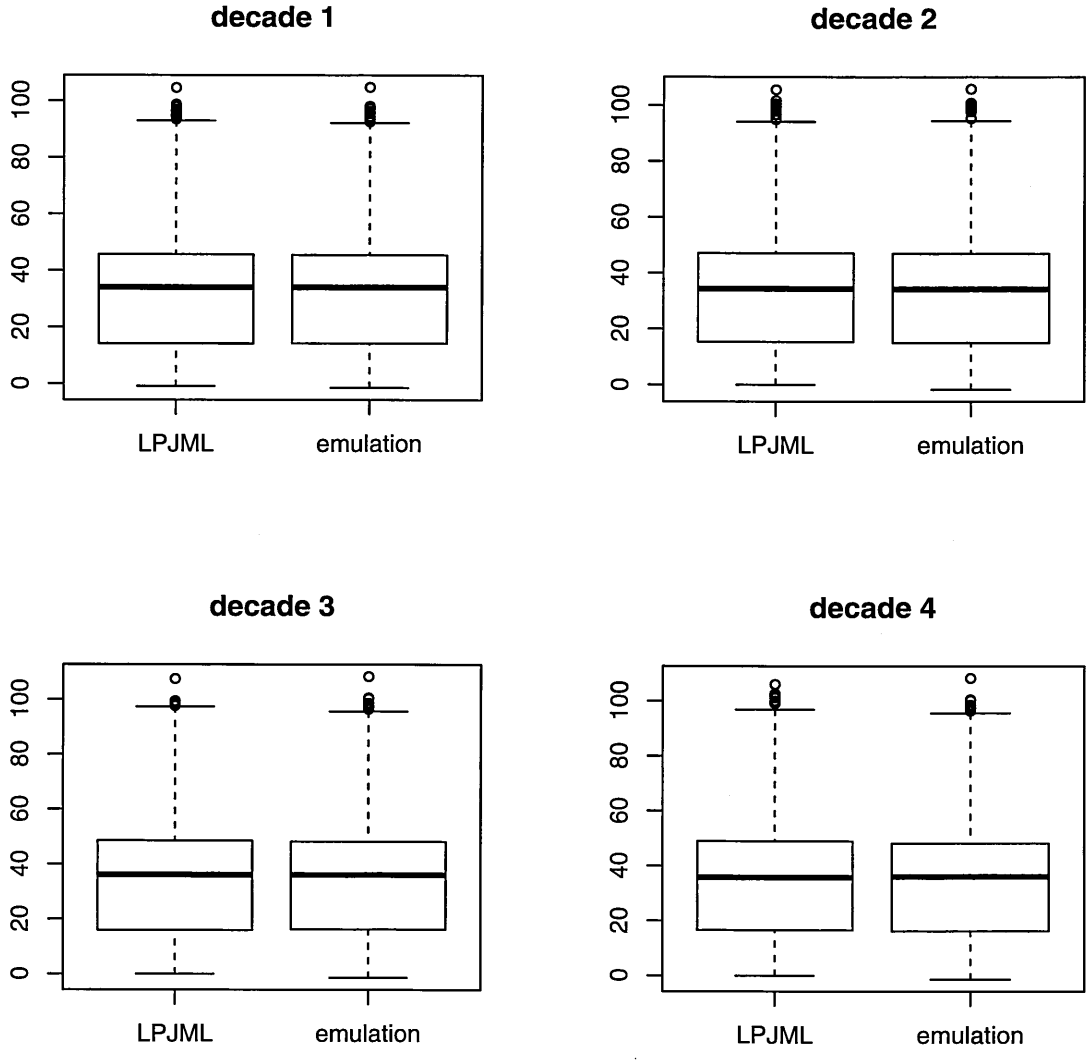


Figure 6.7: Boxplots comparing the observed (simulated by LPJmL) and predicted (emulated) mean decadal NPP in gCm^{-2} for decades (1-4); top-left, top-right, bottom-left and bottom-right plots respectively for RCP3

Figure 6.4(c) is a pairwise plot that shows high correlation between the observed and predicted NPP (correlation of 0.96) for the whole century (aggregated data). This is obtained by adding the previous predicted mean decadal NPP (absolute) to the predicted mean decadal change. It will be used as a guide to try to improve on our emulator. while Figure 6.4(d) is the plot of observed (LPJmL) against predicted (emulated) NPP for (2011-2020) and (2021-2030).

Figure 6.5 shows the mapping of observed and predicted NPP and the two most significant variables for prediction. The fitted model predicted the simulation output well, especially in the tropical climate and high latitude regions. For instance, the predicted and observed NPP change in the upper latitude regions are almost the same.

Figure 6.6 gives the sensitivity index plots for the fitted model of NPP change between decades (1 & 2), (3 & 4), (4 & 5) and (5 & 6). Temperature and wetday frequency were the most significant predictors of NPP change. It is interesting to see that different variables are picked as its significant variables in the different plots. Each decade of data is modelled independently and there is a possibility that the stepwise algorithm picked different combination of terms in the final model.

Figure 6.7 gives the boxplots that compare the distribution of simulated NPP (LPJmL) and the NPP predicted from the emulator for decades (1-4). The median decadal NPP for both the LPJmL and emulator are virtually identical in all four decades (about 40 gCm^{-2}), slightly lower for decades 1 and 2. There are noticeable outliers for both the LPJmL and emulator in each decade.

6.4 Emulation of carbon fluxes using combination of OLS, PCA and WLS: high resolution

We extend the approach of the last section for the 2001-2100 data, where we described the procedure for emulating NPP using OLS on low-resolution data. We would like to emulate NPP, FC and HR at high spatial resolution in order to capture detailed representation of their spatial patterns. High spatial resolution data provide more useful information than low resolution data. Secondly, the modelling in this section is extended into a two-stage approach, where we use the OLS method as in the previous section 6.3 but incorporate an additional step for residual analysis. Residual analysis will enable us to interpolate unexplained variation in the OLS results and improve estimation.

6.4.1 A procedure for statistical emulation

Here we consider an emulation of the change in mean decadal value of carbon fluxes (NPP, FC and HR). We first model the carbon responses using the OLS regression approach. For modelling NPP and FC, we applied precisely the same procedures as in section 6.3 but there is a modification for modelling heterotrophic respiration (HR). The motivation for HR, like NPP and FC, is to understand its dynamics under future climate change. HR is characterized by high spatial variability, and it is usually positively correlated with mean annual temperature (Tuomi et al., 2008; Raich et al., 2002). Our HR modelling will provide a better understanding of the carbon exchange between terrestrial ecosystems and the atmosphere.

Most analyses of HR have considered the change of HR as the temperature increases by 10 °C (Lloyd & Taylor, 1994). A widely used model for the growth of respiration rate as a function of temperature is the exponential function (Tuomi et al., 2008; Suseela et al., 2012). Cox et al. (2000) indicated that temperature, with the availability of sufficient water, will always increase HR. Berridge et al. (2003)

emphasise the dependence of soil respiration on climate and in a recent study, Wang et al. (2014) reinforced the increase in the rate of heterotrophic respiration under global warming. Given the exponential dependence on temperature, we modelled the change in HR between any two successive decades after a logarithm transformation. That is, the dependent variable is

$$\delta = \log(\mathbf{Y_{j+1}}) - \log(\mathbf{Y_j}) = \log\left(\frac{\mathbf{Y_{j+1}}}{\mathbf{Y_j}}\right). \tag{6.2}$$

Where δ is the mean decadal change in HR after a log-transformation and $\mathbf{Y_j}$ is the HR value in j^{th} decade for LPJmL. After transforming the HR data, we then apply equation (6.1) to fit the regression model.

We observed from the OLS results that the correlation between the observed values of LPJML and the predictions from emulator are low and vary between 0.38-0.54 for HR. We tried introducing additional explanatory variables (eg latitude) to improve the models with no substantial improvement. We then performed a residual analysis of the results to look for any visible patterns. The residual plots for HR (see Figures 6.8, 6.9, 6.10) clearly indicated some similarity in spatial patterns both across the GCMs and RCPs. The major motivation for the second stage approach is to improve results. The OLS regression ignores spatial relationships in the dataset, but a grid cell has features that are not all captured by the explanatory variables. The second stage of the analyses allows for a more general multilinear relationship in the residuals. Our approach is related to the technique suggested in O'Hagan (2006), which involves combining a regression with a GP for improved emulator performance. The procedure is described in the next section.

6.4.2 First stage algorithm

The average decadal flux given by LPJmL from 2001-2100 was computed for each carbon flux in each grid cell. We then obtained the change in carbon flux

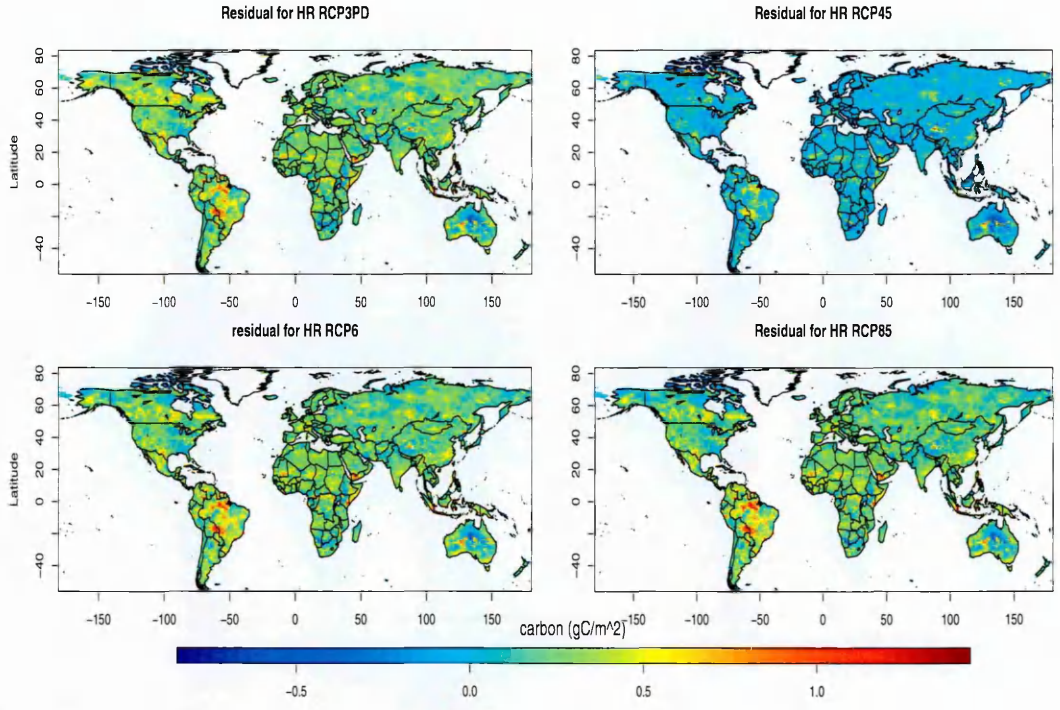


Figure 6.8: Residual map for the mean decadal change in heterotrophic respiration between (2055-2064) and (2065-2074), RCP 6 and RCP8.5 for UKMO-HADGEM1.

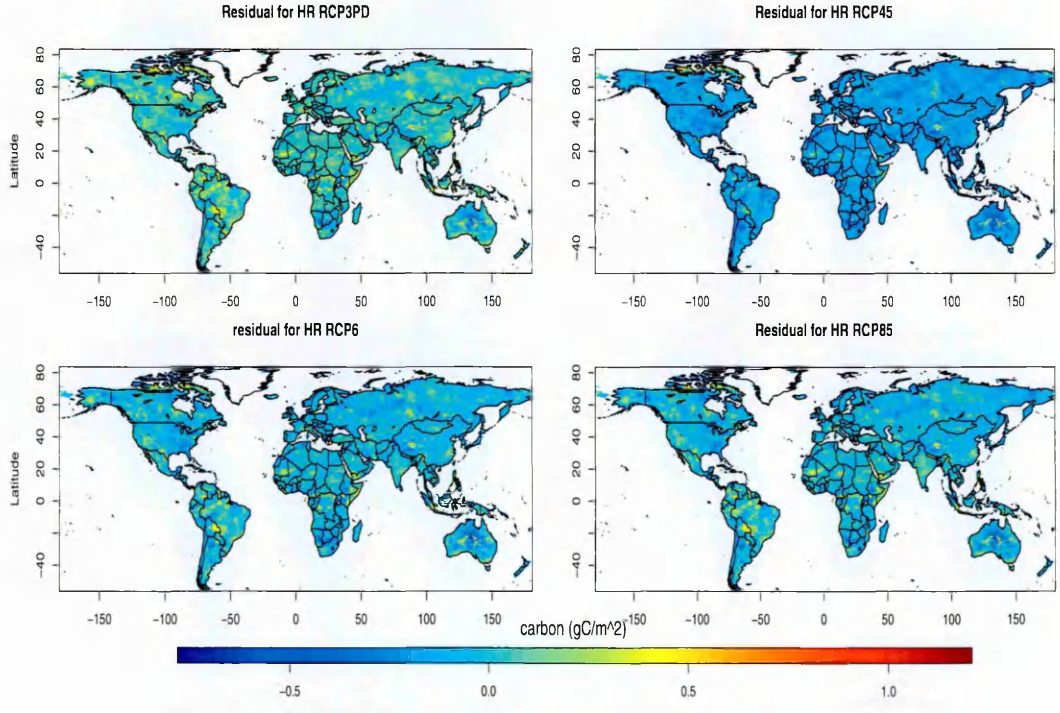


Figure 6.9: Residual map for the mean decadal change in heterotrophic respiration between (2055-2064) and (2065-2074), RCP 6 and RCP8.5 for IPSL-CM4.

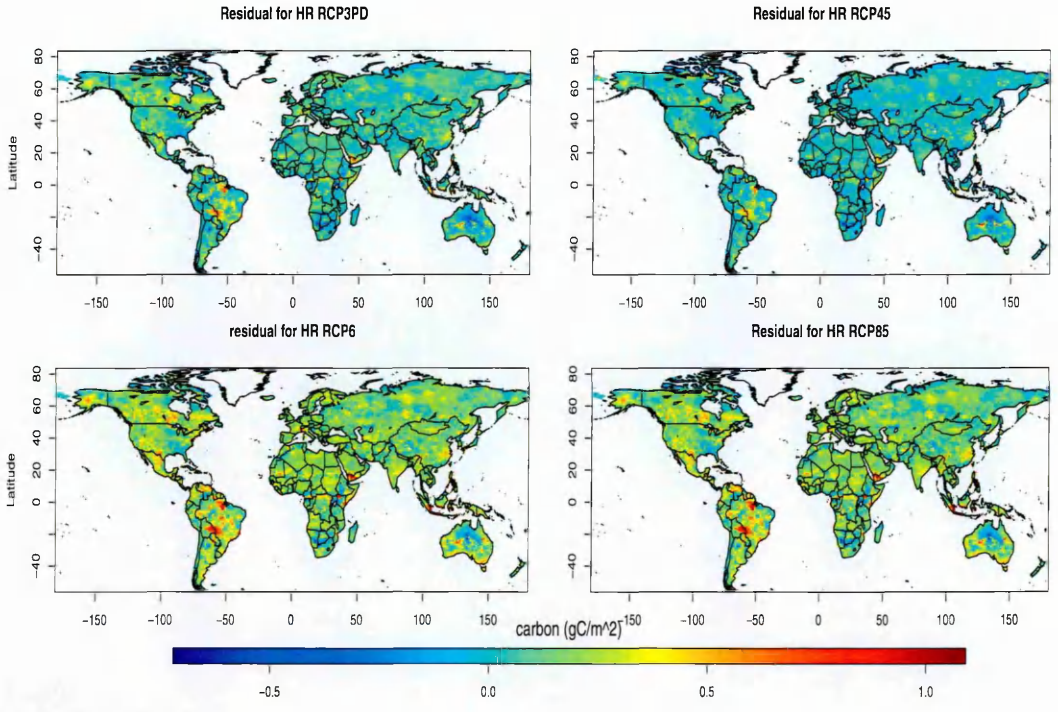


Figure 6.10: Residual map for the mean decadal change in heterotrophic respiration between (2055-2064) and (2065-2074), RCP 6 and RCP8.5 for GISS-MODELEH.

between any two successive decades. We calculated the change in seasonal climate variables for the input variables.

The emulators were built from three GCMs, GISS-MODELER, UKMO-HADGEM1, and IPSL-CM4, four RCPs, giving 12 (3×4) different scenarios. Each scenario has nine time-slices, with each time-slice consisting of 59199 observations. Another GCM, CCSR-MIROC32HI was used for the cross-validation to evaluate the performance of the carbon flux emulators. We use observations from 10000 randomly sampled grid points because the stepwise algorithm could not fit the whole dataset. The 10000 grid points are fixed across the time-slice, RCP, GCM. We fitted a single model to the sampled data for each carbon flux. This procedure enables LPJmL emulators to predict the change in carbon flux for a range of climate forcing scenarios that are not restricted to the RCPs and climate models (GCMs) used to construct them.

An emulator is constructed in two stages. Stage 1 is essentially the same as the OLS (low dimension) emulator described in section 6.3 except that it was

built using the Revolution *R* Enterprise, which has a mechanism for scaling data to handle big computation. We are now using more data because it is high resolution rather than the low resolution. The response variable is the change in carbon flux (NPP, HR and FC) between any two consecutive decades given by LPJmL and is denoted by $\mathbf{y}$. As noted earlier, each combination of RCP and GCM is referred to as a scenario, giving 12 scenarios (3 GCMs, 4 RCPs). Let N denote the number of grid cells for each carbon flux. LPJmL gave values for nine different time slices. Hence $\mathbf{y}$ has $12 \times 9 \times N$ data values, where $\mathbf{y}$ as defined in above. The input variables are listed in Table D.1 of Appendix D excluding LAI and soil.

The explanatory variables can enter the regression as linear or quadratic terms. All two-way interactions were also considered for inclusion. Thus spre, spre^2 and spre.wwet are examples of the potential terms in the regression model.

First stage summary

The following are the steps for building the first stage algorithm of the emulator

(i) Compute the LPJmL average decadal NPP for each time-steps, RCPs, GCMs.

(ii) Randomly sample 10000 grid points from the global data. The sampled grid points are applied across the time-step, RCP, GCM.

(iii) Perform steps (i-ii) on the input variables as well.

(iv) Fit the model using equation 6.1 above with automated stepwise function in Revolution *R*.

(v) Obtain both prediction and error of predictions for all scenarios (time-step, RCP, GCM).

(vi) Repeat steps (i-v) for HR and FC.

6.4.3 Second stage algorithm

The second stage combines the PCA with WLS regression to explain some of the residual variation that is left unexplained by the first stage. In contrast to stage 1, here we form a separate emulator for each time slice. A scenario is defined as any combination of RCP and GCM. In this section, we have 12 scenarios (3 GCMs $\times$ 4 RCPs). The distance between scenario points is the sum of the square differences between the set of known and unknown scenarios scaled by their eigenvalues. This will be used in the computation of estimated weights for the WLS regression. Consider just a single time point for NPP and let $\mathbf{y}_i$ be the vector of changes in NPP given by LPJmL for that combination in the i^{th} known scenario ($i = 1, \ldots, 12$). We let $\tilde{\mathbf{y}}_i$ be the corresponding predictions given by the stage 1 emulator and $\boldsymbol{\varepsilon}_i = \mathbf{y}_i - \tilde{\mathbf{y}}_i$ is the error in prediction. Each $\tilde{\mathbf{y}}_i$ and $\boldsymbol{\varepsilon}_i$ is an $N \times 1$ vector, where N denotes the number of grid cells for the NPP (N=59199). As $\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{12}$ are predictions on *training* data, the values of $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_{12}$ are known.

Given a new vector of predictions, $\tilde{\mathbf{y}}^\star$, from a fresh scenario (any arbitrary RCP/GCM combinations) of NPP where the LPJmLem values are unknown, the aim is to estimate the error of $\tilde{\mathbf{y}}^\star$ from $\mathbf{E} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_{12})$ and $\tilde{\mathbf{Y}} = (\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{12})$. We apply a PC decomposition to the $12 \times N$ matrix $\tilde{\mathbf{Y}}^{\mathrm{T}}$ and select just the first four principal components. The resulting 12×4 matrix $\mathbf{X}_0$ of PC scores given by these four components is then used as explanatory variables for the WLS regression of our residual patterns $\mathbf{E}$. Details are given in the next section.

Our method in this stage is similar to a pattern scaling approach that is commonly used in climate scenario generation. Pattern scaling assumes that given any particular point in space and time, there exists a linear relationship between climate change pattern and global mean temperature with a constant spatial pattern. Here, we allow for a more general multilinear relationship in the residual (c.f. Holden et al. (2014)). As noted earlier, the residual patterns from the OLS results in stage 1 indicated that the residual patterns are relatively

similar across RCP and GCM (see Figures 6.8, 6.9 and 6.10 for HR residual patterns). Hence, for example, if a grid cell has a negative residual in one scenario, then that grid cell is likely to have a negative residual for other scenarios. Stage 2 exploits the similarity between the error patterns across scenarios.

Having obtained the residuals $\mathbf{E}$ by calculating the differences between the emulator predictions and the actual LPJmL values for each scenario, we then interpolate these residual patterns using distance-weighted regression. More weight is assigned to known scenarios that are similar in pattern to the unknown scenario, with similarity determined by the distances from the known scenario points $\tilde{\mathbf{y}}_i, \ldots, \tilde{\mathbf{y}}_{12}$ to the unknown scenario $\tilde{\mathbf{y}}^\star$; closer scenarios are more similar and receive greater weight.

There is a need to take account of the distance between scenario points when modelling the residual. We tried linear, square and cubic distance metrics. We chose a squared distance method scaled by the eigenvalues because it is amongst the best metrics in terms of the proportion of variance it explained (equation 5.72). The same weighting scheme is applied across all grid points. The weights form the non-zero elements of a 12×12 diagonal matrix, $\mathbf{W}$.

A separate weighted regression is performed for each grid cell. For the n^{th} cell, the dependent variable is the n^{th} row of $\mathbf{E}$, $(n = 1, \ldots, N)$ but the data matrix for the explanatory variables $(\mathbf{X}_0)$ and the weight matrix $(\mathbf{W})$ are the same for each grid cell. Thus most of the calculations for estimating regression coefficients need only be performed once, rather than once for each grid cell. From the stage 1 predictions for the new scenario, $\tilde{\mathbf{y}}^\star$, we calculate its PC score $(\tilde{\mathbf{x}}^\star)$ and use the regression equation for the n^{th} grid cell to estimate the error in prediction for that cell. If $\hat{\varepsilon}_n^\star$ is the resulting estimate then we revise the prediction for the cell by adding $\hat{\varepsilon}_n^\star$ to it.

6.4.4 Calculation in stage 2

Predictions given by stage 1 for the 12 training scenarios form the $N \times 12$ matrix $\tilde{\mathbf{Y}} = (\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{12})$. Usually a principal component (PC) analysis would be applied to $\tilde{\mathbf{Y}}$: in the standard data analysis problem, each column of $\tilde{\mathbf{Y}}$ would relate to a different variable and the aim would be to condense the different variables into a few components. Here, each column of $\tilde{\mathbf{Y}}$ corresponds to a different scenario and these different scenarios must retain their identities. However, we want to condense the information given by the N grid points into a few summary statistics. Thus, we want to reduce the number of rows (rather than the number of columns) so a principal component analysis is applied to $\tilde{\mathbf{Y}}^{\mathrm{T}}$. The spectral decomposition theorem gives

$$\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^{\mathrm{T}} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^{\mathrm{T}} \tag{6.3}$$

where $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_N)$ is a diagonal matrix of eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$, and $\mathbf{\Gamma} = (\gamma_1, \ldots, \gamma_N)$ is an $N \times N$ orthogonal matrix, where γ_i is the eigenvector corresponding to λ_i. The rank of $\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^{\mathrm{T}}$ is 12 (or possibly less), so λ_{13}, λ_{14} ... and λ_N each equal 0. The PC transformation from $\tilde{\mathbf{y}} \in \mathbb{R}^N$ to $\tilde{\mathbf{x}} \in \mathbb{R}^{12}$ is given by

$$\tilde{\mathbf{y}} \rightarrow \mathbf{\Gamma}^{\mathrm{T}}\tilde{\mathbf{y}} = \tilde{\mathbf{x}}. \tag{6.4}$$

Put $\tilde{\mathbf{x}}_i = \mathbf{\Gamma}^{\mathrm{T}}\tilde{\mathbf{y}}_i$ for $i = 1, \ldots, 12$ and put $\tilde{\mathbf{x}}^\star = \mathbf{\Gamma}^{\mathrm{T}}\tilde{\mathbf{y}}^\star$, where $\tilde{\mathbf{y}}^\star$ are the stage 1 predictions for the new scenario. Then $(\tilde{\mathbf{y}}^\star - \tilde{\mathbf{y}}_i)^{\mathrm{T}}(\tilde{\mathbf{y}}^\star - \tilde{\mathbf{y}}_i) = (\tilde{\mathbf{x}}^\star - \tilde{\mathbf{x}}_i)^{\mathrm{T}}(\tilde{\mathbf{x}}^\star - \tilde{\mathbf{x}}_i)$. We wish to give more importance to the eigenvectors that correspond to large eigenvalues, so we define the distance between $\tilde{\mathbf{y}}^\star$ and $\tilde{\mathbf{y}}_i$ as

$$d_i = \left[\sum_{j=1}^{N} \lambda_j (x_j^\star - x_{ij})^2\right]^{1/2} = \left[\sum_{j=1}^{12} \lambda_j (x_j^\star - x_{ij})^2\right]^{1/2}, \tag{6.5}$$

where $x_j^\star$ and x_{ij} are the j^{th} components of $\tilde{\mathbf{x}}^\star$ and $\tilde{\mathbf{x}}_i$, respectively.

In forming regression equations, we use only the first four principal components; for every carbon flux/time slice combination, these explained at least 95%

of the variation in $\tilde{\mathbf{Y}}$. Let $\hat{\mathbf{\Gamma}}$ be the first four columns of $\mathbf{\Gamma}$, so $\hat{\mathbf{\Gamma}} = (\gamma_1, \gamma_2, \gamma_3, \gamma_4)$. We put $\hat{\mathbf{X}} = \tilde{\mathbf{Y}}^{\mathrm{T}}\hat{\mathbf{\Gamma}}$.

Let $\mathbf{y}_i$ be the $N \times 1$ vector of observations given by LPJmL for the i^{th} scenario of the training set ($i = 1, \ldots, 12$); $\tilde{\mathbf{y}}_i$ is the corresponding vector of prediction given by stage 1, and $\boldsymbol{\varepsilon}_i = \mathbf{y}_i - \tilde{\mathbf{y}}_i$ gives the error in the stage 1 predictions. A weighted least squares (WLS) regression is used to estimate $\boldsymbol{\varepsilon}^\star$, the corresponding error in $\tilde{\mathbf{y}}^\star$. A separate regression equation is determined for each component of $\boldsymbol{\varepsilon}^\star$. We put $\mathbf{E} = (\boldsymbol{\varepsilon}_i, \ldots, \boldsymbol{\varepsilon}_{12})$ so for the j^{th} regression (the j^{th} grid cell) the values of the dependent variable are $\tilde{\boldsymbol{\varepsilon}}_j^{\mathrm{T}}$, where $\tilde{\boldsymbol{\varepsilon}}_j^{\mathrm{T}}$ is the j^{th} row of $\mathbf{E}$.

If any distance d_i is close to 0, then weighted regression is unnecessary as the scenario with the zero-distance gets a weight of infinity, and the errors for that scenario are taken as the errors in $\tilde{\mathbf{y}}^\star$. Sometimes the distance equals 0 for more than one scenario and then the errors for those scenarios are averaged. Specifically, if Q denotes the set of integers such that $d_i = 0$ for $i \in Q$, then the vector of estimated error for the new scenario is $\hat{\boldsymbol{\varepsilon}}^\star = \sum_{i \in Q} \boldsymbol{\varepsilon}_i$. For non-zero d_i, the weights w_i ($i = 1, \ldots, 12$) and weight matrix $\mathbf{W}$ are defined as

$$w_i = \frac{(1/d_i^2)}{\sum_1^{12}(1/d_i^2)} \qquad \text{and} \qquad \mathbf{W} = \begin{pmatrix} w_1 & 0 & \ldots & 0 \\ 0 & w_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & w_{12} \end{pmatrix}.$$

We take one grid point at a time and form a separate regression equation for that grid point. The data for one of these regressions is the (12×1) vector of responses $\tilde{\boldsymbol{\varepsilon}}_j$ (the errors at that grid point) and the (12×4) matrix $\hat{\mathbf{X}}$, which holds the values taken by the explanatory variables. So as to include a constant term in the regression model, we put $\mathbf{X}_0 = \hat{\mathbf{X}}$. The weighted linear regression uses $\mathbf{W}$ as the weight-matrix.

For the n^{th} grid point, the regression model is $E(\boldsymbol{\varepsilon}_n \,|\, \mathbf{x}_0) = \boldsymbol{\beta}_n^{\mathrm{T}}\mathbf{x}_0$ where $\boldsymbol{\varepsilon}_n$ is a random variable whose observed value is the i^{th} component of $\tilde{\boldsymbol{\varepsilon}}_n$ when $\mathbf{x}_0^{\mathrm{T}}$ is

the i^{th} row of $\mathbf{X}_0$. The WLS estimate of $\boldsymbol{\beta}_n$ is

$$\hat{\boldsymbol{\beta}}_n = (\mathbf{X}_0^{\mathrm{T}}\mathbf{W}\mathbf{X}_0)^{-1}\mathbf{X}_0^{\mathrm{T}}\mathbf{W}\tilde{\boldsymbol{\varepsilon}}_n. \tag{A.4}$$

For the new scenario, put $\mathbf{x}_0^\star = \hat{\boldsymbol{\Gamma}}^{\mathrm{T}}\tilde{\mathbf{y}}^\star$. Then the estimate of the prediction error in $\tilde{\mathbf{y}}^\star$ at the n^{th} grid point is $\varepsilon_n^\star = \hat{\boldsymbol{\beta}}_n^{\mathrm{T}}\mathbf{x}_0^\star$, which (using (A.4)) can be written as

$$\hat{\varepsilon}_j^\star = \tilde{\boldsymbol{\varepsilon}}_j^{\mathrm{T}}\mathbf{W}\mathbf{X}_0(\mathbf{X}_0^{\mathrm{T}}\mathbf{W}\mathbf{X}_0)^{-1}\mathbf{x}_0^\star. \tag{6.6}$$

Equation (6.6) estimates the prediction error separately for each grid cell. To combine the calculations for all grid cells into a single step, let $\hat{\boldsymbol{\varepsilon}}^\star = (\hat{\varepsilon}_1^\star, \ldots, \hat{\varepsilon}_N^\star)^{\mathrm{T}}$. Then the equation for estimating all the residuals in $\tilde{\mathbf{y}}^\star$ is given as

$$\hat{\boldsymbol{\varepsilon}}^\star = \mathbf{E}^{\mathrm{T}}\mathbf{W}\mathbf{X}_0(\mathbf{X}_0^{\mathrm{T}}\mathbf{W}\mathbf{X}_0)^{-1}\mathbf{x}_0^\star. \tag{6.7}$$

The similarity of our method to pattern scaling stems from this equation. The vector $\hat{\boldsymbol{\varepsilon}}^\star$ is the estimated error pattern for the new scenario and $\boldsymbol{\varepsilon}_i$ is the (known) error pattern for the i^{th} training scenario. If we put $\mathbf{W}\mathbf{X}_0(\mathbf{X}_0^{\mathrm{T}}\mathbf{W}\mathbf{X}_0)^{-1}\mathbf{x}_0^\star = (\alpha_1, \ldots, \alpha_{12})^{\mathrm{T}}$, then (6.7) may be written as

$$\hat{\boldsymbol{\varepsilon}}^\star = \sum_{i=1}^{12} \alpha_i \boldsymbol{\varepsilon}_i,$$

the estimated error pattern for the new scenario is simply a linear combination of the error patterns of the training scenarios.

To avoid extrapolation, we bound the n^{th} component of $\hat{\boldsymbol{\varepsilon}}^\star$ to be within the range of $\tilde{\boldsymbol{\varepsilon}}_n$. Let $\hat{\boldsymbol{\varepsilon}}^\#$ denote the resulting vector when a component of $\hat{\boldsymbol{\varepsilon}}^\star$ is set equal to any bound it exceeds. We take $\hat{\mathbf{y}}^\star = \tilde{\mathbf{y}}^\star + \hat{\boldsymbol{\varepsilon}}^\#$ as the emulated value of $\mathbf{y}$ for the new scenario.

Second stage summary

The following is a summary of the calculations in Stage 2.

(i) Perform a principal components analysis of $\tilde{\mathbf{Y}}^{\mathrm{T}}\tilde{\mathbf{Y}}$. The non-zero eigenvalues are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{12}$ and the corresponding eigenvectors are $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{12}$. Put $\tilde{\mathbf{x}}^\star = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{12})^{\mathrm{T}}\tilde{\mathbf{y}}^\star$ and $\tilde{\mathbf{x}}_i = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{12})^{\mathrm{T}}\tilde{\mathbf{y}}_i$ for $i = 1, \ldots, 12$.

(ii) Denote the j^{th} components of $\tilde{\mathbf{x}}^\star$ and $\tilde{\mathbf{x}}_i$ by $x_j^\star$ and x_{ij}, respectively. Then $w_1, \ldots, w_{12}$ are the non-zero elements of the diagonal matrix $\mathbf{W}$, where $w_i = (1/d_i^2)/\{\sum_1^{12}(1/d_i^2)\}$, with $d_i^2 = \sum_{j=1}^{12} \lambda_j (x_j^\star - x_{ij})^2$.

(iii) The explanatory variables for the WLS regression are constructed from the first four eigenvectors of $\boldsymbol{\Gamma}$. We put $\hat{\boldsymbol{\Gamma}} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, \boldsymbol{\gamma}_4)$ and $\mathbf{X}_0 = \tilde{\mathbf{Y}}^{\mathrm{T}}\hat{\boldsymbol{\Gamma}}$.

(iv) Weighted least squares gives $\hat{\boldsymbol{\beta}}_n = (\mathbf{X}_0^{\mathrm{T}}\mathbf{W}\mathbf{X}_0)^{-1}\mathbf{X}_0^{\mathrm{T}}\mathbf{W}\tilde{\boldsymbol{\varepsilon}}_n$ as the vector of regression coefficients for the n^{th} grid cell, where $\tilde{\boldsymbol{\varepsilon}}_n^{\mathrm{T}}$ is the n^{th} row of $\mathbf{E}$.

(v) The estimated error for the n^{th} grid cell is $\hat{\varepsilon}_n^\star = \hat{\boldsymbol{\beta}}_n^{\mathrm{T}}\mathbf{x}_0^\star$, where $\mathbf{x}_0^\star = \hat{\boldsymbol{\Gamma}}^{\mathrm{T}}\tilde{\mathbf{y}}^\star$.

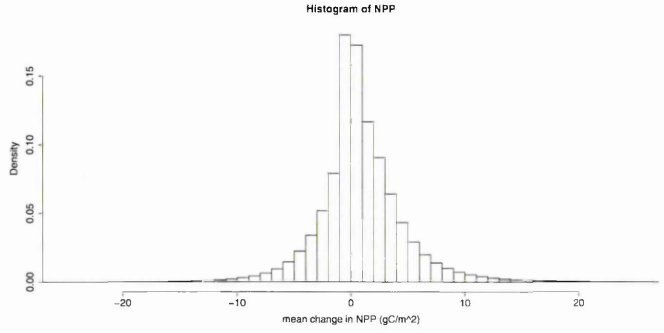
(vi) In order to avoid unusual values and prevent too much extrapolation from known residuals, we compare $\hat{\varepsilon}_n^\star$, with each component of $\tilde{\boldsymbol{\varepsilon}}_n$. Let $(\varepsilon_n^{min}, \varepsilon_n^{max})$ denote the range of these components. If $\hat{\varepsilon}_n^\star$ is outside this range, we set it equal to the range's nearer endpoint. Thus the revised estimate of y for the n^{th} grid point is $\tilde{y}_n^\star + \hat{\varepsilon}_n^\#$, where $\tilde{y}_n^\star$ is the n^{th} component of $\tilde{\mathbf{y}}^\star$ and

$$\hat{\varepsilon}_n^\# = \begin{cases} \varepsilon_n^{min} & \text{if } \hat{\varepsilon}_n^\star < \varepsilon_n^{min} \\ \varepsilon_n^\star & \text{if } \varepsilon_n^{min} \leq \hat{\varepsilon}_n^\star \leq \varepsilon_n^{max} \\ \varepsilon_n^{max} & \text{if } \hat{\varepsilon}_n^\star > \varepsilon_n^{max}. \end{cases} \tag{6.8}$$

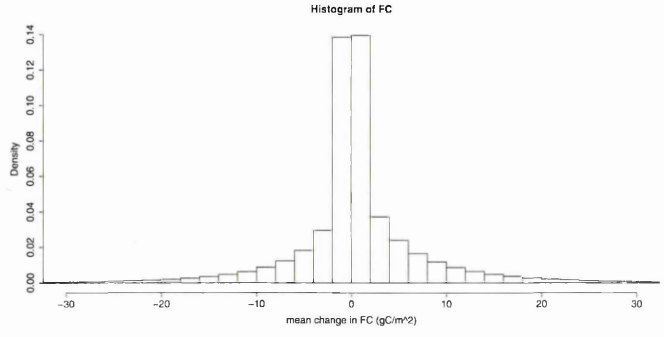
Our methods described above are flexible to apply. We present the results from the carbon fluxes emulator after performing the residual analysis of the second stage.

6.5 Results

Before fitting the OLS regression to the 2001-2100 dataset, we first plot the histograms of our response data in Figure 6.11 to show the distribution and summarise the data. We can see from Figures 6.11 that the distributions are fairly symmetric and normal, about 95% of the NPP data are within -10 to 15 gC/m^2, for fire carbon that varies between -20 to 20 gC/m^2. Similarly, Figure 6.12 which is for respiration after logarithm transformation, the values range from -0.4 to 0.5.



(a)



(b)

Figure 6.11: Histogram of mean decadal changes in NPP and FC respectively (gC/m^2) for all time points, RCPs and 3GCMs.

Figures 6.13, 6.13 and 6.13 are pairwise diagnostic plots for NPP, FC and HR. They compare the performance of the emulator among RCPs for the periods between (2055-2064) and (2065-2074). We can see that the correlation between the emulator and LPJmL is high for NPP in Figure 6.13 with the 45° passing through the center of the cloud. LPJmL outputs are relatively well predicted in all the four RCPs and similarly for HR in Figure 6.14. Fire carbon is also fairly

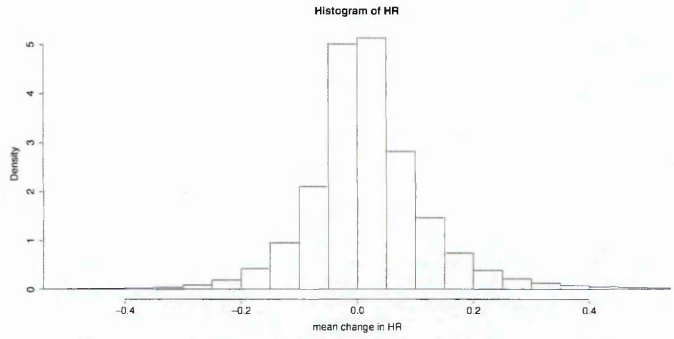


Figure 6.12: Histogram for the mean decadal change in HR for all time points, RCPs and 3GCMs. NOTE: these values have been log-transformed.

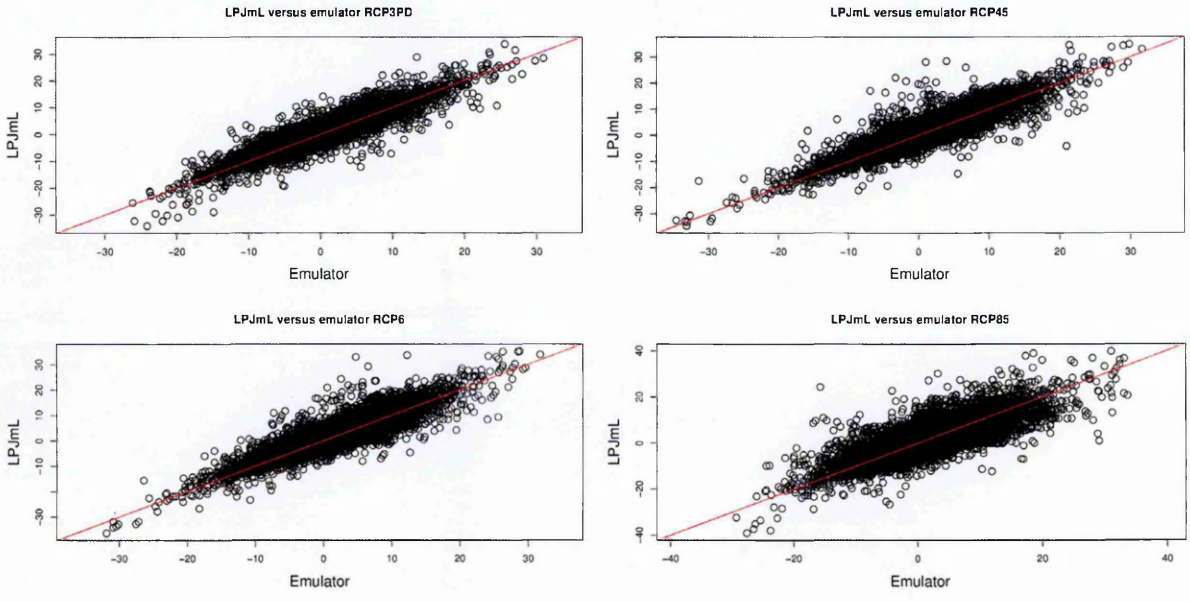


Figure 6.13: LPJmL versus emulator prediction for the mean decadal change in NPP (gC/m^2) between (2055-2064) and (2065-2074), for the four RCPs and CCSR-MIROC32HI.

well predicted. There is a relatively good agreement between the LPJmL values and emulator predictions.

Table 6.3 indicates the performance of the three carbon emulators across the time point (decade). The values are the proportion of variance explained by the emulator for the cross-validated data. The values are decreasing with time especially for NPP, and this is expected because we modelled change in carbon fluxes between two consecutive decades thereby accumulating the emulator error. The proportion is also relatively decreasing with time for FC except at the 1^{st} decadal change that has 57% variance. There is no visible pattern for the respiration.

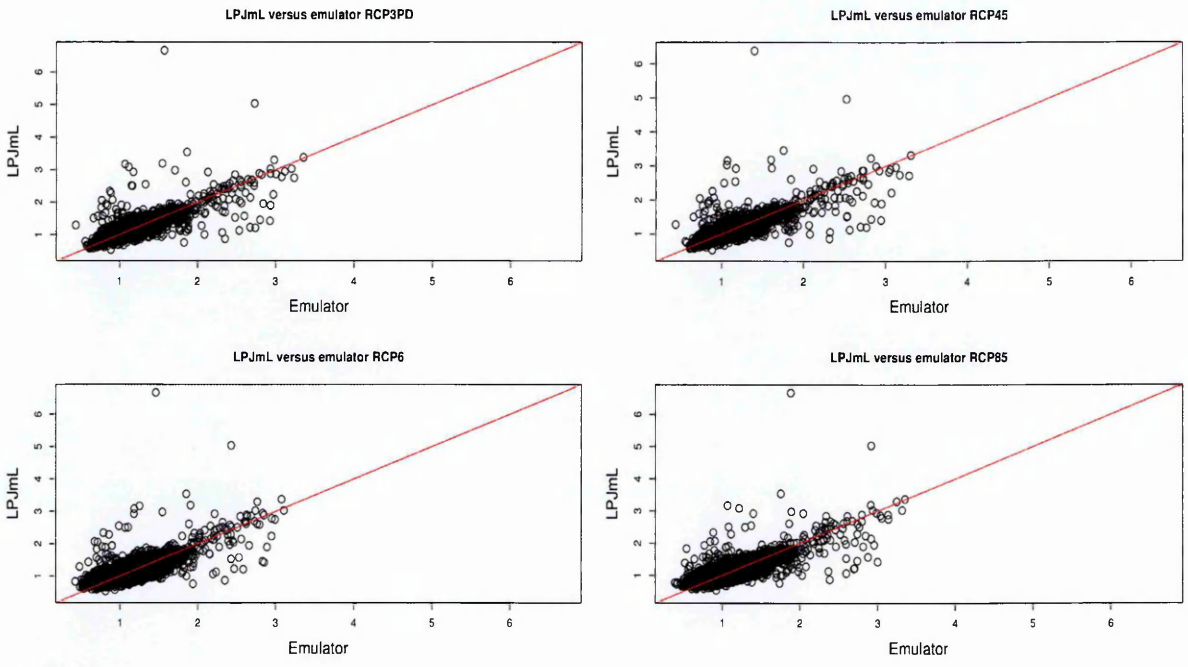


Figure 6.14: LPJmL versus emulator prediction for the mean decadal change in heterotrophic respiration (back-transformed from logarithm) between (2055-2064) and (2065-2074), for the four RCPs and CCSR-MIROC32HI.

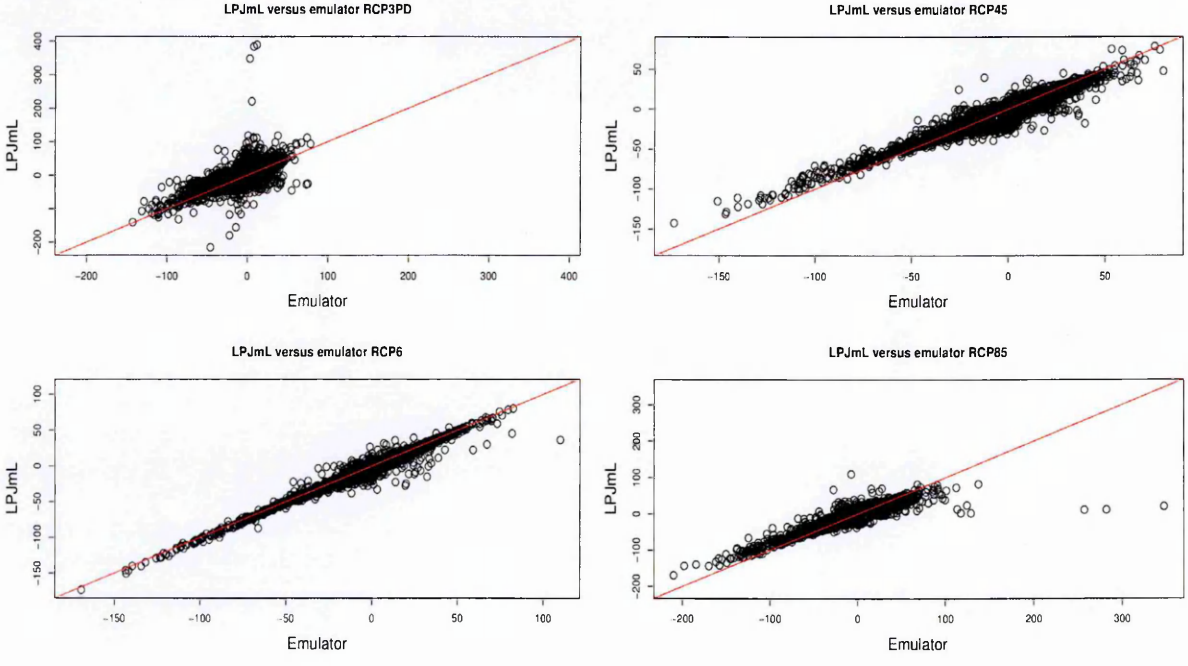


Figure 6.15: LPJmL versus emulator prediction for the mean decadal change in fire carbon (gC/m^2) between (2035-2044) and (2045-2054), for the four RCPs and CCSR-MIROC32HI.

NPP emulator seems to estimate LPJmL values much better than for HR and FC emulator. The percentage of variance explained varies between $66 - 92\%$ for

NPP, between $58-84\%$ for HR and between 45-90% for the fire carbon emulator. Table 6.4 provides the overall performance of the emulator for each RCP. We observed that the emulator performs less well under RCP 8.5 than for other RCPs. This may be because RCP 8.5 is an extreme and so there may be climate states that are not well sampled in the training data. The last column of Table 6.4 gives the overall proportion of variance explained by these emulation.

Table 6.3: Table of cross-validated proportion of variance ρ showing the performance of the emulators for NPP, HR and FC for each time point from CCSR-MIROC32HI.

decade j/carbon	1	2	3	4	5	6	7	8
NPP	0.92	0.87	0.81	0.77	0.71	0.75	0.70	0.66
HR	0.84	0.76	0.70	0.58	0.67	0.69	0.69	0.60
FC	0.57	0.90	0.76	0.60	0.59	0.63	0.52	0.45

Table 6.4: Table of cross-validated proportion of variance ρ showing the overall performance of the emulators for NPP, HR and FC for all RCPs and time slices from CCSR-MIROC32HI.

RCP/carbon	RCP3PD	RCP4.5	RCP6	RCP8.5	**Overall**
NPP	0.85	0.78	0.78	0.68	**0.77**
HR	0.76	0.74	0.66	0.60	**0.70**
FC	0.73	0.72	0.73	0.63	**0.61**

Figures 6.16, 6.17 and 6.18 are the spatial maps comparing LPJmL with emulator predictions for NPP, HR and FC. The results are for the period (2055-2064) and (2065-2074), RCP 6 and 8.5. LPJmL and the emulators agree relatively well in most of the regions for both RCPs, although the NPP emulator under RCP 8.5 seems to under-predict the fluxes in Russia and Europe. Large carbon changes are observed in some places in Brazil, Russia and USA, while there are relatively small changes estimated in subtropical regions for both RCPs. The reduction in carbon fluxes in these regions can be attributed to rising temperature and cloudiness. LPJmL is over-estimated by the emulator in every region in Figure 6.17. There are some large extreme changes in a few areas under RCP 8.5. Figure 6.18, comparing the fire carbon, changes are over-estimated by the emulator for RCP6. RCP8.5 is also less well predicted. The potential for wide occurrence of a

large amount of carbon due to fire outbreak is relatively high. Most regions have large increases in carbon loss due to fire disturbance, especially for RCP 8.5.

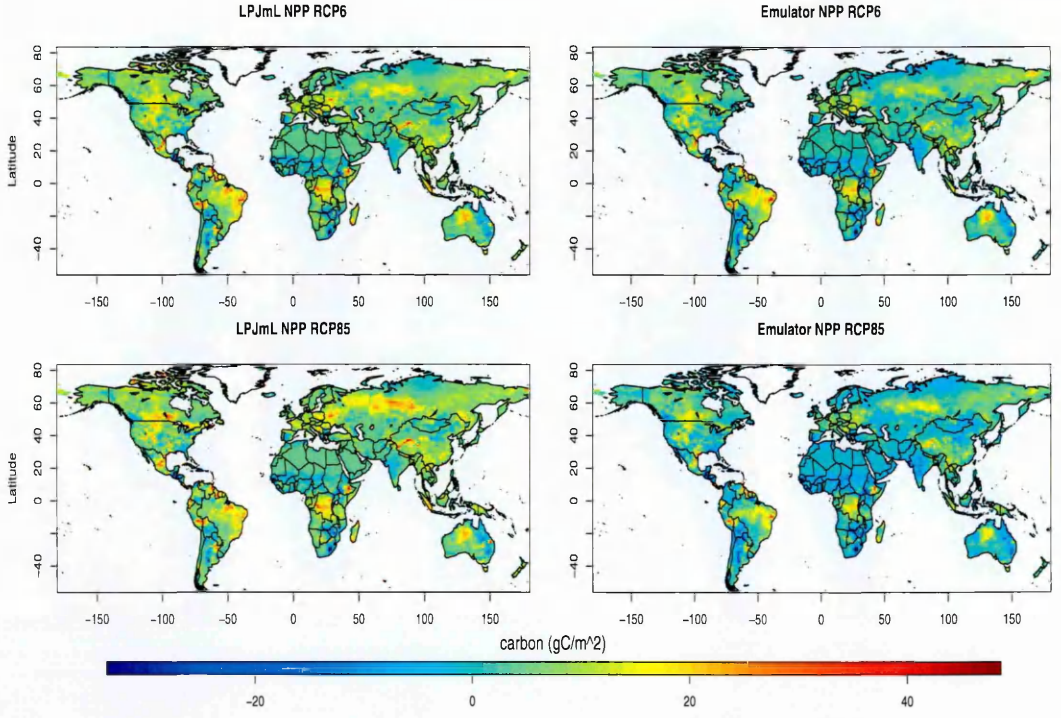


Figure 6.16: Cross-validated spatial map for the mean decadal change in NPP (gC/m^2) between (2055-2064) and (2065-2074), RCP 6 and RCP8.5 for CCSR-MIROC32HI.

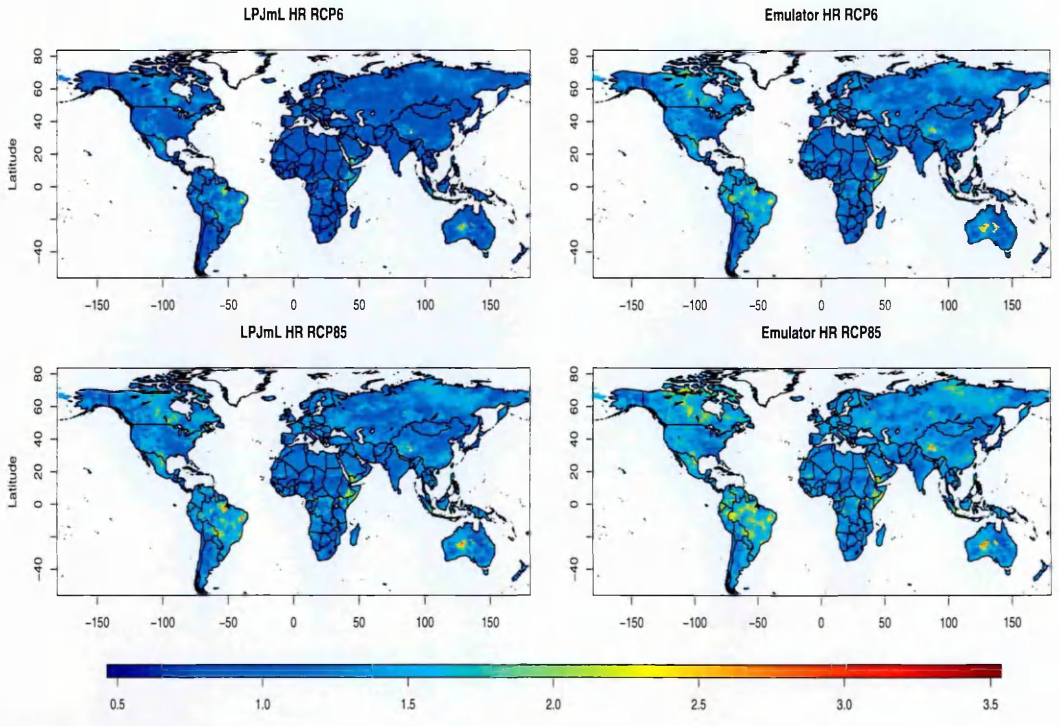


Figure 6.17: Cross-validated spatial map for the mean decadal change in logarithm of heterotrophic respiration (back-transformed from logarithm) between (2055-2064) and (2065-2074), RCP 6 and RCP8.5 for CCSR-MIROC32HI

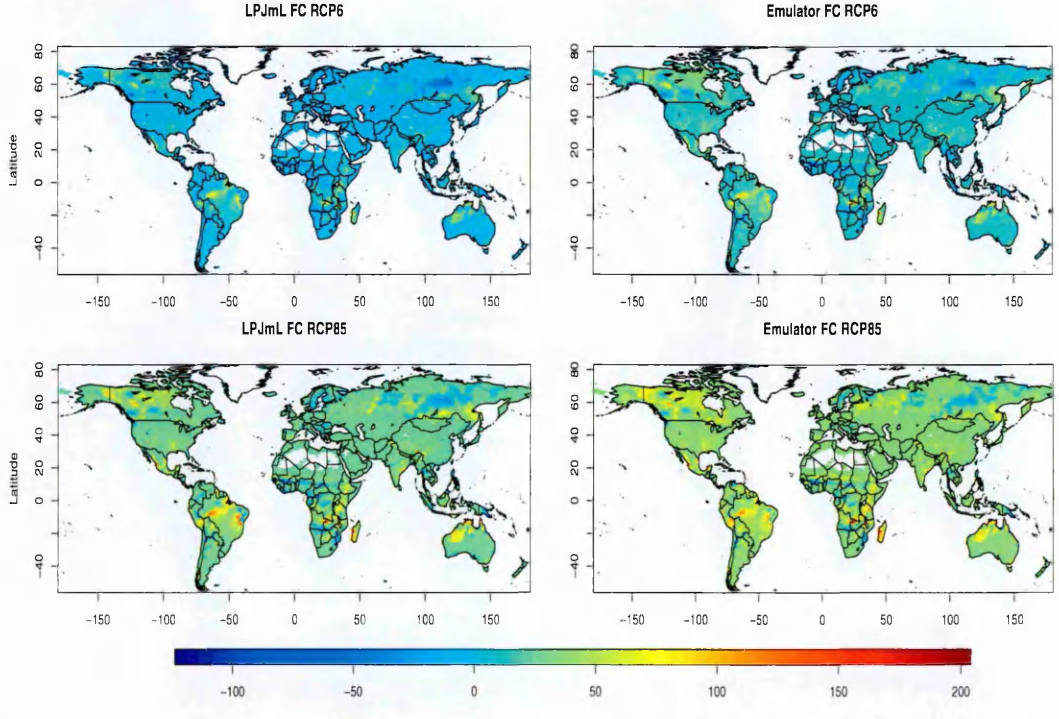


Figure 6.18: Cross-validated spatial map for the mean decadal change in fire carbon (gC/m^2) between (2035-2044) and (2045-2054), RCP 6 and RCP8.5 for CCSR-MIROC32HI

6.6 Conclusion

There have been a large number of studies that have investigated carbon flux response under climate change, and various approaches have been considered. In this chapter, we have presented a regression method for predicting terrestrial biospheric response of net primary productivity, heterotrophic respiration and fire carbon to climate change and anthropogenic CO_2 emission. We examined a series of models for estimating mean decadal change in NPP, FC and HR for the 1901-2000 and 2001-2100 datasets.

We analysed both datasets by examining the relationship between NPP, FC, HR and the explanatory variables. The explanatory variables include seasonal climate data (surface temperature, precipitation, wet day frequency, cloud cover, diurnal temperature range, surface temperature minimum, surface temperature maximum) and CO_2 concentration. We fitted several quadratic models to relate the terrestrial biospheric response of each of NPP, HR and FC to climate variables. We tried avoiding over-fitting in our models by limiting our analysis to linear or quadratic terms and used term selection. We used proportion of variance explained ρ to test the model performance and AIC and BIC to select variables. In addition, we performed a sensitivity analysis to identify important environmental variables that will determine the level of carbon fluxes in the future.

We emulated the mean decadal NPP, both absolute and change. We predicted mean decadal change in NPP, FC, HR for each successive decade from (1901-2000) and mean decadal change for the period (2001-2010). The literature on climate change and CO_2 impact of terrestrial biosphere shows that the terrestrial biosphere is sensitive to climate change and CO_2 emissions. The results of our analyses concurred with this finding. The results from this chapter provided useful insight and direction for the remaining work in this thesis. The next chapter will deal with emulation of potential crop yields from LPJmL.

Chapter 7

The emulation of crop yields

This chapter deals with procedures and techniques for the emulation of crop yield data. In addition to having both low and high-resolution emulators for crop yield, we also have rainfed and irrigated crop emulators. In addition to the form of emulator described in Chapter 6 we also build an emulator that used censored regression. These four CFTs were selected for emulation: temperate cereal, rice, maize and oil crops (sunflower, soybeans, rapeseed and groundnut). The crops are chosen because they are widely grown across the globe. Cereal, rice and maize are very prominent staple food crops and provide over 50% of all calories consumed by the world. Oil crops are used domestically as a vegetable oil and also forms ingredients used in manufacturing products. The emulation procedures are described below.

7.1 Emulation of crop yield using censored regression

We consider the use of censored regression to model the following agricultural crops: cereal, rice, maize and oil. Oil is modelled as the average of soybean, groundnut, rapeseed and sunflower in this section. Analysis here is motivated by the presence of many zeros in the crop simulation data. Censored regression

is used when the data on the response variable are limited, or it is difficult to observe the full response variable. Observations clustered on threshold values or at the endpoints of a permitted range are classified as censored data.

Ordinary least squares is a procedure usually used to compute the best fit line to data while censored regression is applicable when the recorded data on the dependent variable cuts off outside a certain range, typically with multiple observations at the minimum or maximum points of that range. When the data are censored, variations in the observed dependent variable will downplay the effects of the regressor on the dependent variable. We already know that using least squares regression to analyse this crop data will make the estimated coefficients biased towards zero (Kenneth & James, 2001; Cleves et al., 2008).

Because the crop yield data have a large proportion of observations at their minimum point (zeros), the data are not observable on the entire range (in the sense that negative yields are not meaningful), which causes the estimate of both the variance and mean to be biased. Some locations might be a "long way below 0", so that a moderate improvement in the climate would still leave them with a yield of 0. Other locations might be "quite close to 0", so that they would have a positive yield if the climate improves slightly. Our model must distinguish between these situations. So in this section, we use censored regression to analyse the crop data.

7.1.1 Crop yields: low-resolution results

The LPJmL crop-yield data were based on simulations for 59199 grid cells on 0.5° by 0.5° as discussed under subsection 4.3.2 of Chapter 4. In this section, we shall focus only on rainfed crops and one management levels using three GCMs, IPSL-CM4, GISS-MODELER and CCSR-MIROC32HI to build the emulator and UKMO-HADGEM1 for cross-validation.

We reduced the dimension of the dataset by aggregating to 2° by 2° resolution. We then obtained mean decadal yield for cereal, rice and maize for each decade in

the data range (2001-2100). We used change in seasonal climate, initial climate, baseline yield, CO_2 and change in CO_2 as emulator inputs.

We computed change in seasonal climate for each successive decade (precipitation, wet-day frequency, cloud cover and temperature). We used initial seasonal climate (previous mean decadal climate variables) as explanatory variables. We formed a matrix of initial crop yields (previous decadal mean yield) which provides historical information for the prediction. We used equation (6.1) to model the data with inclusion of CO_2 and change in CO_2 as additional inputs to the model, see Table 7.1.

Table 7.1: The emulator's input variables for censored regression.

Variables	Full names
	2001-2100 data
scld/weld/spcld/acld	Change in mean cloud cover in summer/winter/spring/autumn
spre/wpre/sppre/apre	Change in mean precipitation in summer/winter/spring/autumn
stmp/wtmp/sptmp/atmp	Change in mean temperature in summer/winter/spring/autumn
swet/wwet/spwet/awet	Change in mean wet day frequency in summer/winter/spring/autumn
iscld/iweld/ispcld/iacld	Initial (baseline) mean cloud cover in summer/winter/spring/autumn
ispre/iwpre/isppre/iapre	Initial mean precipitation in summer/winter/spring/autumn
istmp/iwtmp/isptmp/iatmp	Initial mean temperature in summer/winter/spring/autumn
iswet/iwwet/ispwet/iawet	Initial wet day frequency in summer/winter/spring/autumn
CO_2 and CCO_2	Initial (baseline) mean CO_2 and change in mean CO_2

In the northern hemisphere, *summer* = {June July August}, *winter* = {December January February}, *spring* = {March April May} and *autumn* = {September October November}; obvious changes are made for the southern hemisphere.

The emulators are built using the *survival* package in (R, 2013). We also combined this approach with a stepwise regression algorithm from the *MASS* package of (R, 2013) to reduce the number of terms in the model. We fitted a quadratic model to the decadal mean crop yield changes. We use equation (6.1), where $\mathbf{y}$ in this case is the decadal mean change in crop yield. The crop yield data are censored observations, p is the number of parameters for estimation and $x_1, \ldots, x_p$ are p independent variables.

We used censored regression for the emulation of crop yields. Censored regression algorithm often takes time to fit a large dataset (i.e., it is computational demanding). We need to reduce the number of observations in order to speed-up the algorithm, so we used just five decades of data. We obtained five time-slices

(2001-2060) of mean decadal crop yield change for rice, temperate cereal, maize and oil. Oil is modelled as the average of soybean, groundnut, rapeseed and sunflower in this section. We fitted a single model that combined three different climate models (GCMs). Another dataset from a fourth climate model was left out for the cross-validation.

The reason for incorporating multiple GCM inputs and cross-validating against unused input data, is to ensure that the emulator is representing the underlying relationship between climate variables and vegetation dynamics relatively well, rather than simply reproducing a spatial pattern of vegetation change on the basis of a spatial pattern of climate change that is characteristic of the models used. Since different climate models will produce quite different patterns of climate change in some regions, testing the prediction with a different climate model input, not used to build the emulator, provides some confidence that underlying relationships are being represented. Multiple GCMs will capture the variability in the climate models and allow for one source of uncertainty in future projections of vegetation dynamics.

We slightly modified the stepwise regression approach used for carbon flux emulation in Chapter 6, to accommodate three major steps in fitting our final model in this section. We first fitted the regression starting from the 1^{st} order terms. We used all the relevant explanatory variables in the model. Secondly, we chose BIC for the backward variable elimination procedure to reduce the linear model terms. This step will remove less relevant variables from our results. It would also prepare the algorithm for more rigorous variable selection. Thirdly, we fitted the second stepwise model to only the variables selected in the previous stage. But here, we allow the model terms to grow to a higher degree of quadratic terms (including two-way interactions). The variables are added step by step until the algorithm converges using AIC as the criteria for inclusion. The BIC selection criterion that we used is more penalized (selected fewer variables) than the AIC method. We further improved the model by allowing the highly

significant variables to grow to cubic terms.

The essence of performing the stepwise procedure this way is to ensure that in the final model we have fewer cases where interaction terms are included before their main effects, which are much likely under the stepwise procedure in Chapter 6. We summarized the censoring procedure below.

1. Obtain the indicator of the censoring for the emulator of the change in yield, using the following rule. Let $Y_{1,i}$ and $Y_{2,i}$ be the vector of actual yield given by LPJmL for any two consecutive decades, with $i = 1, \ldots, 2257$. Then the change in yield and corresponding censored indicator are given respectively as

$$\Delta Y = Y_{2,i} - Y_{1,i}$$

$$\delta_i = \begin{cases} 1, & \text{if } Y_{2,i} > 0 \ \& \ Y_{1,i} > 0 \quad \text{(uncensored)} \\ 2, & \text{if } Y_{2,i} > 0 \ \& \ Y_{1,i} < 0 \quad \text{(right-censored)} \\ 3, & \text{if } Y_{2,i} < 0 \ \& \ Y_{1,i} > 0 \quad \text{(left-censored)} \\ 0, & \text{if } Y_{2,i} < 0 \ \& \ Y_{1,i} < 0 \quad \text{(truncated)} \end{cases} \tag{7.1}$$

2. Perform the cross-validation by obtaining the predictions for each of the change in yield emulators.

Here we provide plots of the results comparing LPJmL values with those of our built emulator. The emulators are able to reproduce the global pattern of crop yield change in response to climate change and CO_2 emissions. Table 7.2 gives a comparison of the performance of the least squares and censored regression methods. The results show that the censored regression out-performed the OLS method in term of percentage of variability of response explained by the model. The results support the use of censored regression to model our data. The spatial maps below compare the change in mean decadal yield from LPJmL with the built emulators for the cereal, rice and maize (Figures 7.1, 7.2, 7.3). The emulators do an excellent job of reproducing the spatial pattern of crop yield given by LPJmL.

Table 7.2: Comparison of correlation of OLS and censored regression for the crop emulators on 2 by 2 degree resolution. There are 33855 observations in our data. This corresponds to 5 decadal changes of data period (2005-2065); each decadal change has 2257 observations and there are 3 GCMs. % of observations censored are the proportion of observations at 0 out of the total 33855 observations.

Crop	% of observations censored	ρ (OLS)	ρ (Censored regression)
cereal	42.8	0 .771	0.798
rice	75.6	0.536	0.667
maize	40.8	0 .688	0.710

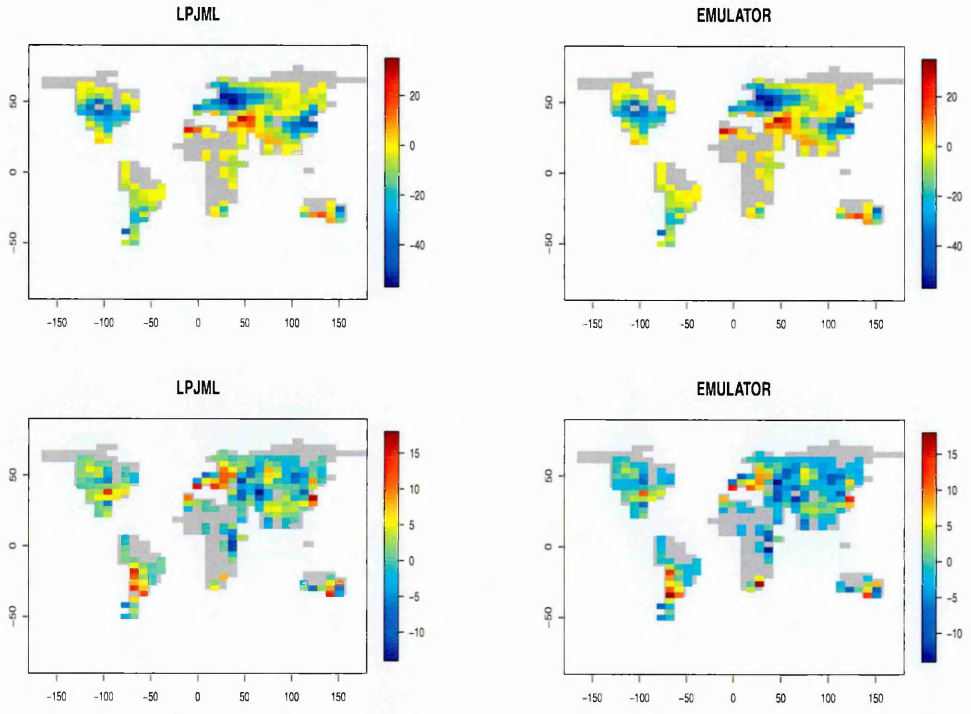


Figure 7.1: Change in mean decadal yield for cereal using a censored regression, between (2001-2010) & (2011-2020)(top) and (2011-2020) & (2021-2030) (bottom) (top-left) LPJmL (top-right) emulator (left) LPJmL (right) emulator from UKMO-HADGEM1.

7.1.2 Crop yields: high-resolution results

In this section, we move back to the original half degree scale of the data as we did under carbon flux in Chapter 6. The procedure here differs a little from the previous section 7.1.1 in that here we are constructing two different emulators, one for the actual yield and the other for a change in yield. We have earlier seen in section 7.1.1 that the censored regression can capture information on the non-responses (zero observations) in our model. The approach treats the zeros

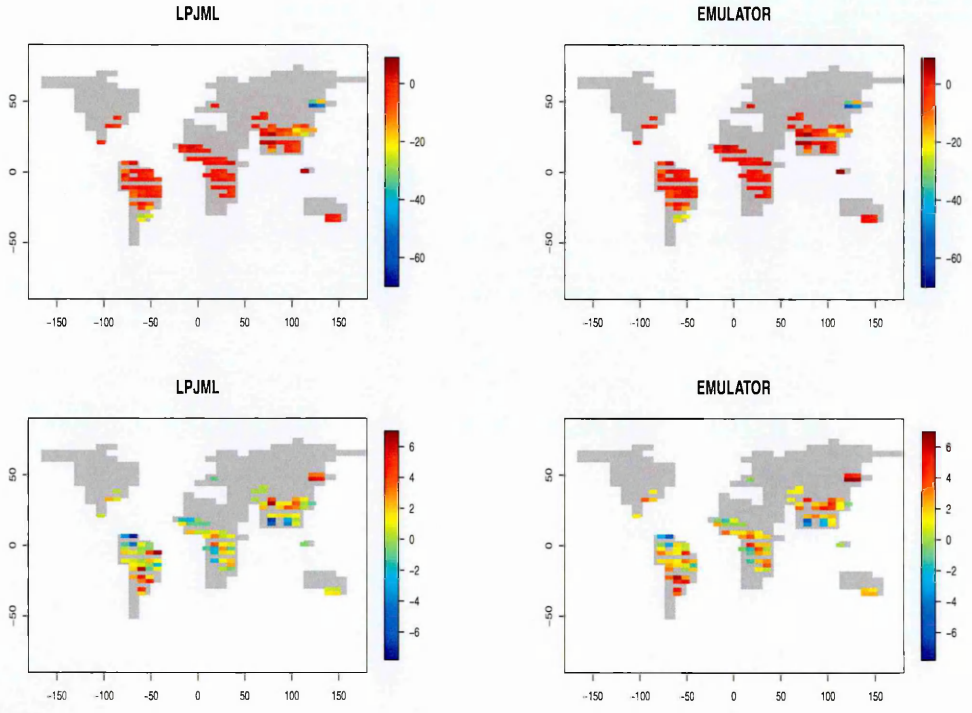


Figure 7.2: Change in mean decadal yield for rice using a censored regression, between (2001-2010) & (2011-2020)(top) and (2011-2020) & (2021-2030) (bottom) (top-left) LPJmL (top-right) emulator (left) LPJmL (right) emulator from UKMO-HADGEM1.

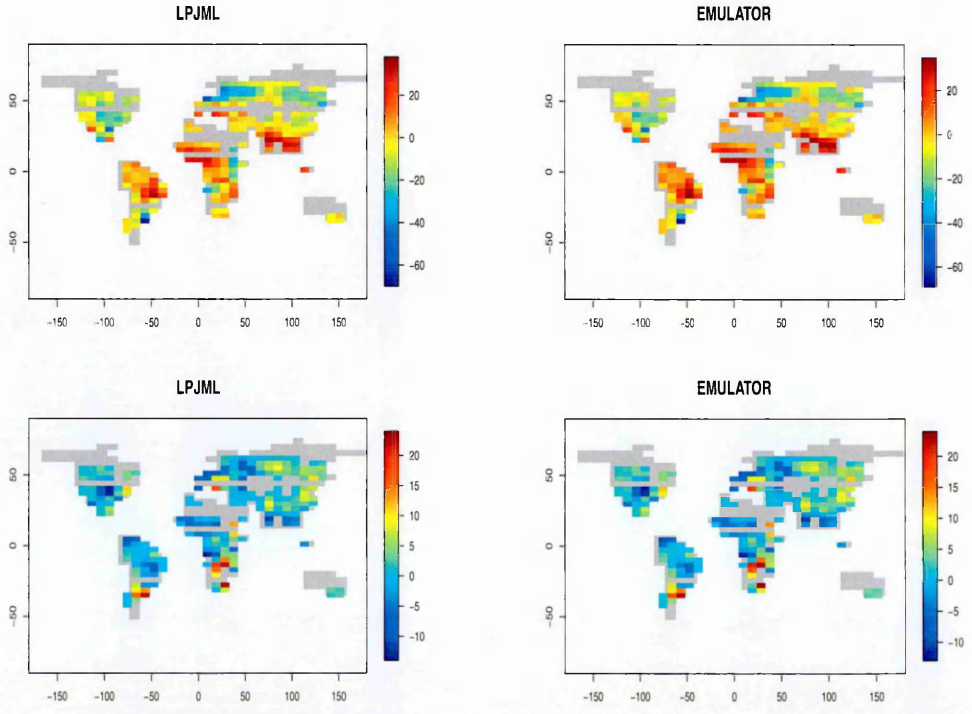


Figure 7.3: Change in mean decadal yield for maize using a censored regression, between (2001-2010) & (2011-2020)(top) and (2011-2020) & (2021-2030) (bottom) (top-left) LPJmL (top-right) emulator (left) LPJmL (right) emulator from UKMO-HADGEM1.

as censored observation (not real observed values). As crop yield is a continuous variable taking only the value zero (where the plant does not grow) and some positive values, our data are left censored. The change in crop yield is a doubly censored variable as there is a possibility of both positive and negative observations as well as zeros. In other words, we have both left and right-censored data present in the change in crop yield (while the actual yield is only left-censored).

We first built emulators for the actual yield for cereal, rice and maize. We made some predictions from these actual yield emulators. We then used these predictions to determine the spatial points on the globe that will be censored. For instance, those points where the actual yield emulators produce positive predictions will be right-censored. This will be left-censored for corresponding negative predictions, and where there is no change of sign in the predictions, those points are left uncensored. We also completely truncated (removed) any points where the crop are not currently growing (grid cell that has zero in both decades).

Unlike the last subsection (7.1.1), here we constructed the emulators for the change in crop yield directly from the actual yield emulators as follows.

1. Obtain the indicator of the censoring for the emulator of the actual yield. Suppose that $Y = Y_1, \ldots, Y_i$, is a vector of actual yield given by LPJmL. Then the indicator δ_i of actual yield emulator censoring, is defined as follows.

$$\delta_i = \begin{cases} 1, & \text{if } Y_i > 0 \text{ (uncensored)} \\ 0, & \text{if } Y_i = 0 \text{ (censored)} \end{cases}$$

2. Construct an actual yield emulator for each crop.

3. Compute the crop predictions for each of the actual yield emulators.

4. Obtain the indicator of the censoring for the emulator of the change in yield, using the following rule. Let $Y_{1,i}$ and $Y_{2,i}$ be the vector of predictions given by actual yield emulator for any two consecutive decades, with $i =$

$1, \ldots, 59199$. Then the change in yield and corresponding censored indicator are given respectively as

$$\Delta Y = Y_{2,i} - Y_{1,i}$$

$$\delta_i = \begin{cases} 1, & \text{if } Y_{2,i} > 0 \ \& \ Y_{1,i} > 0 \quad \text{(uncensored)} \\ 2, & \text{if } Y_{2,i} > 0 \ \& \ Y_{1,i} < 0 \quad \text{(left-censored)} \\ 3, & \text{if } Y_{2,i} < 0 \ \& \ Y_{1,i} > 0 \quad \text{(right-censored)} \\ 0, & \text{if } Y_{2,i} < 0 \ \& \ Y_{1,i} < 0 \quad \text{(truncated)} \end{cases} \tag{7.2}$$

Results of censored regression are only provided for maize under RCP3 and 8.5, Figure 7.4. We note that the censored regression emulator gives the best results for maize. It produces fairly good results in most places but does not capture the extreme values as well, extreme data are under-estimated in RCP 3 and over-predicted in RCP 8.5. a limitation of this approach is that censored regression uses a numerical method that often required a long time for the estimation of the likelihood parameters, as we discussed under section 5.3.2 in Chapter 5. In addition, the results are not much better than for the corresponding OLS results. Besides, the algorithm could not handle all the data at once. This prompted a move back to the two-stage method that was used for carbon flux in Chapter 6.

7.2 Emulation of crop yields using combination of OLS, PCA, WLS

The emulator is essentially the same as the two-stage emulator described in Chapter 6, but some details differ and are described in this section.

7.2.1 A procedure for statistical emulation

The LPJmL crop-yield data were based on simulations for 59199 grid cells on 0.5° by 0.5°, but here we only consider those cells where crops actually grow in

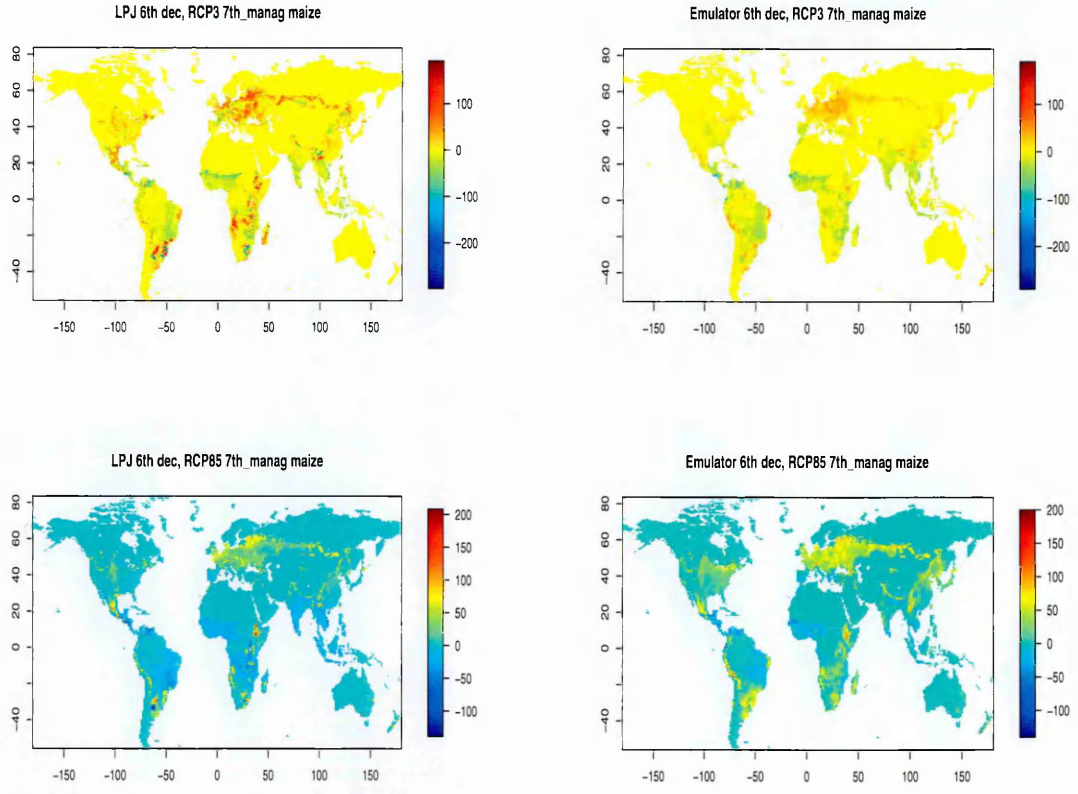


Figure 7.4: Spatial map for the mean decadal change in yield (gC/m^2) for maize as these give the best result using censored regression between (2055-2064) and (2065-2074), RCP 3 and RCP8.5 for management 7 under UKMO-HADGEM1.

the simulation (the other cells are truncated). We built separate emulators for the rainfed and irrigated crops.

Groundnut is categorised as a separate group in this section. Therefore, the following five CFTs were selected for emulation; temperate cereal, rice, maize, groundnut and oil crop. The oil crop is a maximum yield among soybean, sunflower and rapeseed which is different from previous definitions. Groundnut crop is emulated separately in this section because it is not appropriate to group the four oil crops together. Functionally rapeseed is quite different to groundnut and therefore might respond quite differently to climate change as we mentioned in Chapter 2. In addition, we observed from the sensitivity results under section 7.1 (not shown) that oil is sensitive to winter temperature which might have been driven by the vernalization response in rapeseed, whereas there is no obvious

reason why groundnut would be sensitive to winter temperatures.

The average decadal yield given by LPJmL from 2005-2095 was computed for each crop and each scenario. We then obtained the change in yields relative to the baseline values. Baseline values are the average decadal yield of LPJmL outputs and climate data for the period 2005-2014, and are used as a reference for computing the relative change in yields and climate. We calculated the change in seasonal climate variables for the input variables listed in Table D.1. Baseline yields, baseline seasonal climate and change in seasonal climate variables are used as input to the emulator. CO_2, latitude, soil type and LAI_{max} are also included as additional inputs.

The emulators were constructed in two stages (see figure 7.5). We built a single emulator for the two CO_2 fertilization levels ("on" and "off"), but treat irrigated and rainfed crops separately. This allows the emulator to be flexible in predicting yield changes for any level of CO_2.

The emulators were built from two GCMs with relatively moderate equilibrium climate sensitivity, CCCMA-CGCM31 and CCSR-MIROC32HI, four RCPs, and two simulation categories (with and without CO_2 fertilization effect), giving 16 ($2 \times 4 \times 2$) different scenarios. Each scenario has seven crop management levels and eight time-slices, with each time-slice consisting of 59199 observations. After removing the zero observations from the data, there are 16649, 6913, 19139, 7100 and 8427 valid grid points per scenario for rainfed temperate cereal, rice, maize, groundnut and oil, respectively, and 8963, 7325, 9146, 2386 and 2731 for irrigated crops. In order to evaluate the performance of LPJmLem another five GCMs, UKMO-HADGEM1, GISS-MODELER, GISS-MODELEH, IPSL-CM4 and CCSR-MIROC32MED were used for cross-validation purposes.

We used all the simulation data for irrigated oil and groundnut emulators because each of these crops has less than 5000 valid grid cells in each scenario. For other crops, we use observations from 5000 randomly sampled grid points because the stepwise algorithm could not fit the whole dataset. The 5000 grid

points are fixed across the time-slice, RCP, GCM and management levels as well as simulation categories and similarly in each of 37 input variables (see Table D.1) in Appendix. We then fitted a single model to the sampled data for each crop-yield when crops are rainfed. This procedure was carried out for each of the five crops and repeated for the case where crops are irrigated. We use cross-validation to examine emulator performance, testing the emulators on five climate models that played no part in their construction.

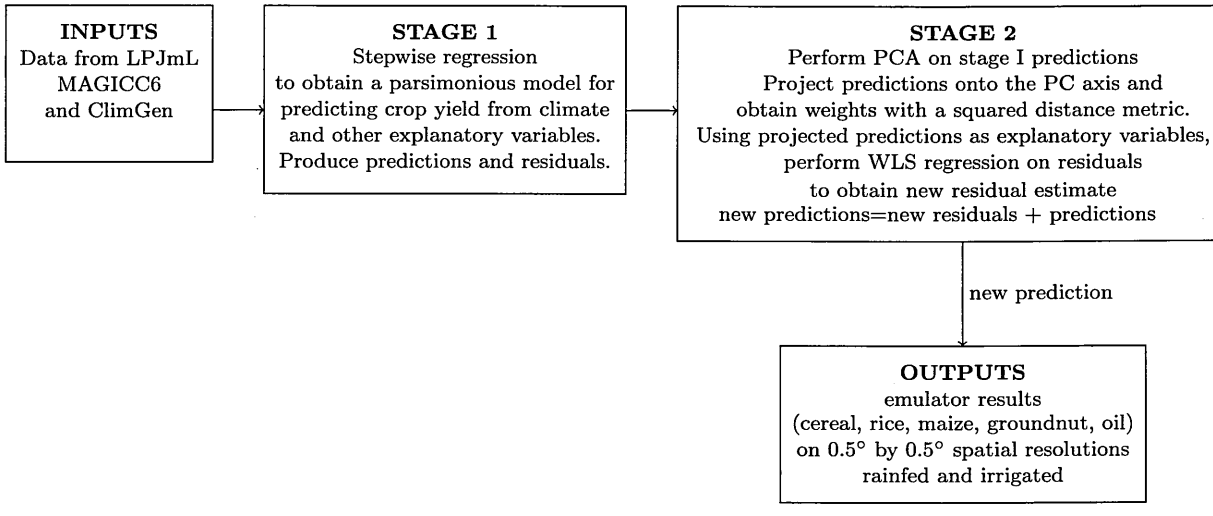


Figure 7.5: Stages for emulator construction.

7.2.2 First stage algorithm

In the stepwise regression of stage 1 the response variable is the change in yield given by LPJmL. As noted earlier, each combination of RCP, GCM and CO_2 fertilization level is referred to as a scenario. Here this gives 16 scenarios rather than the 12 scenarios used in Chapter 6. Letting N denote the number of grid cells for the combination of crop and irrigation regime of current interest, for each scenario we have $7 \times 8 \times N$ data as LPJmL gave values for seven management levels and eight different time slices. Hence $\mathbf{y}$ has $16 \times 7 \times 8 \times N$ data values. The explanatory variables are listed in Table D.1 and now include soil and LAI.

An integer with values between 1 and 7 was used to represent the LAI_{max} parameter and this formed a factor variable in the regression analysis. The other

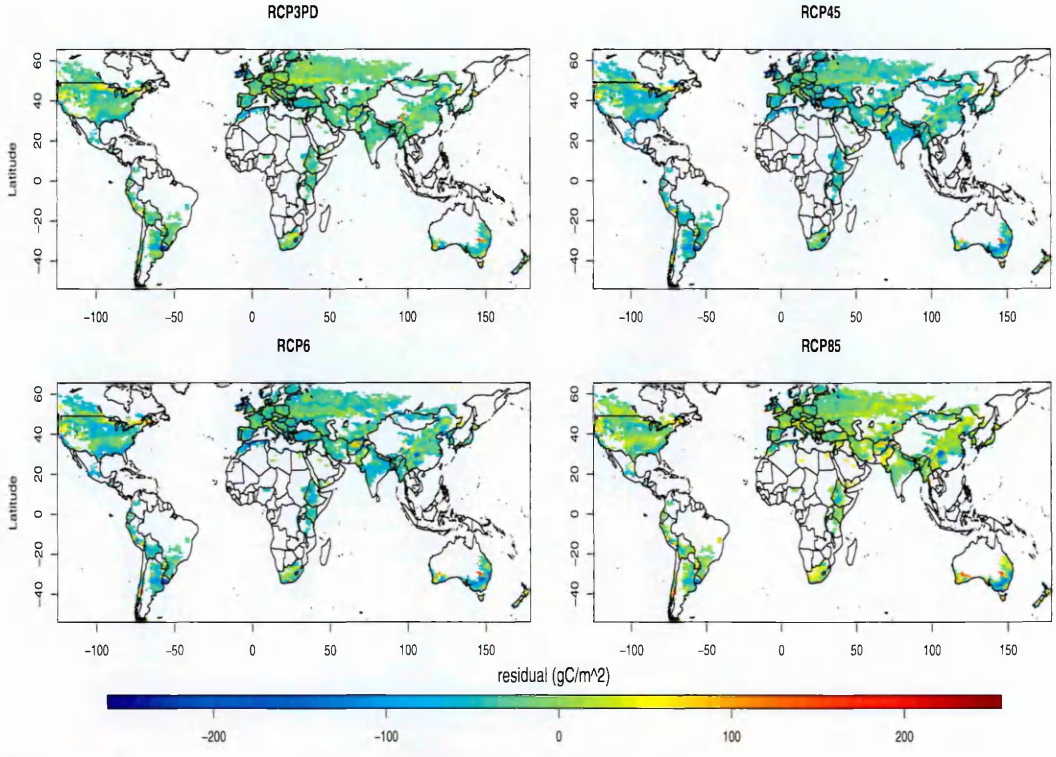
explanatory variables can enter the regression as linear or quadratic terms. As before, all two-way interactions were also considered for inclusion. The regression model has the form given by equation (6.1).

7.2.3 Second stage

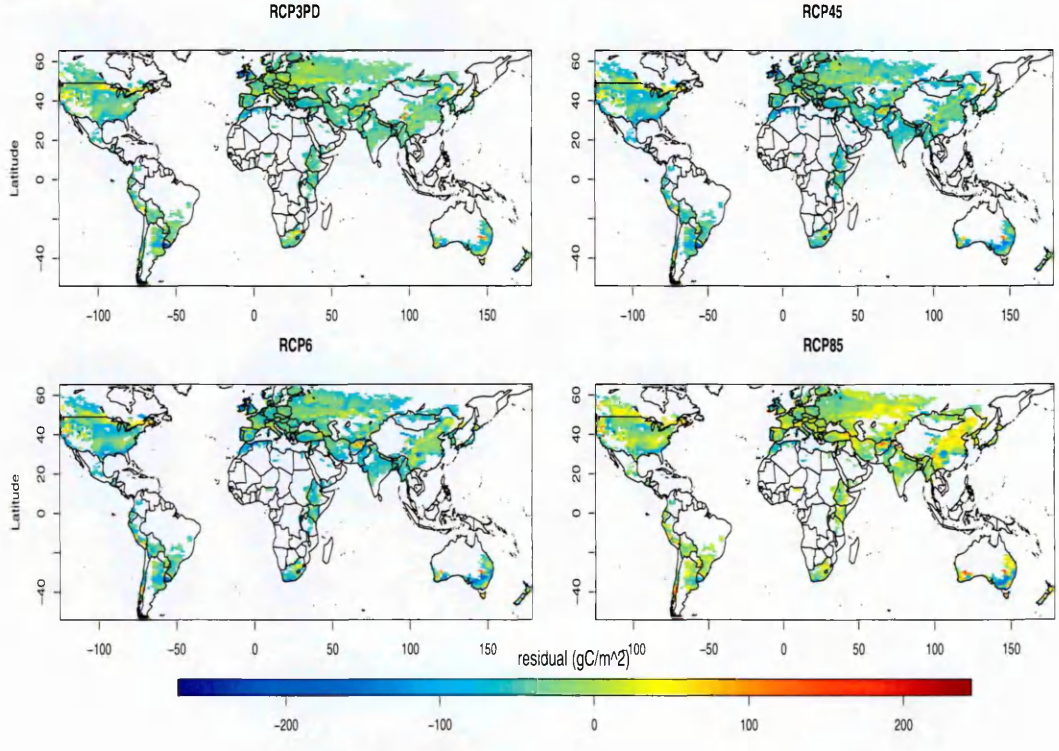
We formed a separate emulator for each combination of crop, irrigation regime, management level and time slice. Let $\mathbf{y}_i$ be the vector of changes in yield given by LPJmL for that combination in the i^{th} scenario ($i = 1, \ldots, 16$). The $\tilde{\mathbf{y}}_i$ is the corresponding predictions given by the stage 1 emulator and $\boldsymbol{\varepsilon}_i = \mathbf{y}_i - \tilde{\mathbf{y}}_i$ is the error in prediction. Each $\tilde{\mathbf{y}}_i$ and $\boldsymbol{\varepsilon}_i$ is an $N \times 1$ vector, where N denotes the number of grid cells for that crop/irrigation regime.

As when emulating carbon fluxes, the residual patterns from the OLS results in stage 1 for crop yields indicated that the patterns are relatively similar across RCP and GCM (see Figures 7.6 and 7.7). Figures 7.6 and 7.7 are two maps of residuals from stage 1 for rainfed cereal. They show a marked degree of similarity across scenarios. In Chapter 6, we only used linear, square and cubic distance metrics. In this section, we considered additional distance metrics namely spherical, exponential, Gaussian and three forms of Matern metric. Details of these metrics are given in Chapter 5 section 5.6.5.

A variogram result is also given in Chapter 7 section 7.3.2 that illustrates the need to take account of the distance between scenario points when modelling stage 1 prediction error. Performances of the metrics under validation data were compared. We chose a squared distance method scaled by the eigenvalues as in Chapter 6 because it is amongst the best metrics in terms of the proportion of variance it explained. We then fitted a separate weighted regression for each grid cell. The remaining procedure is the same as in subsection (6.4.4) of Chapter 6 except that here we used 16 different scenarios. Thus where subscripts have an upper limit of 12, that limit must now be changed to 16. For clarity, a summary of the second stage is given again with that change.

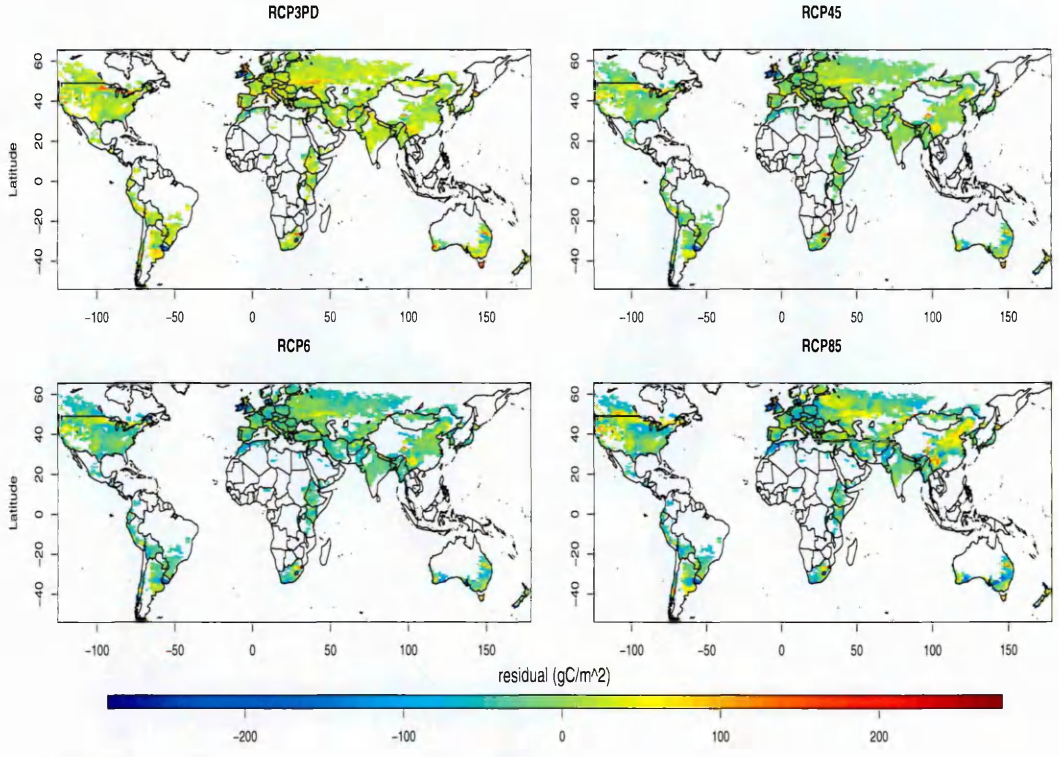


(a) With CO_2 fertilization effect

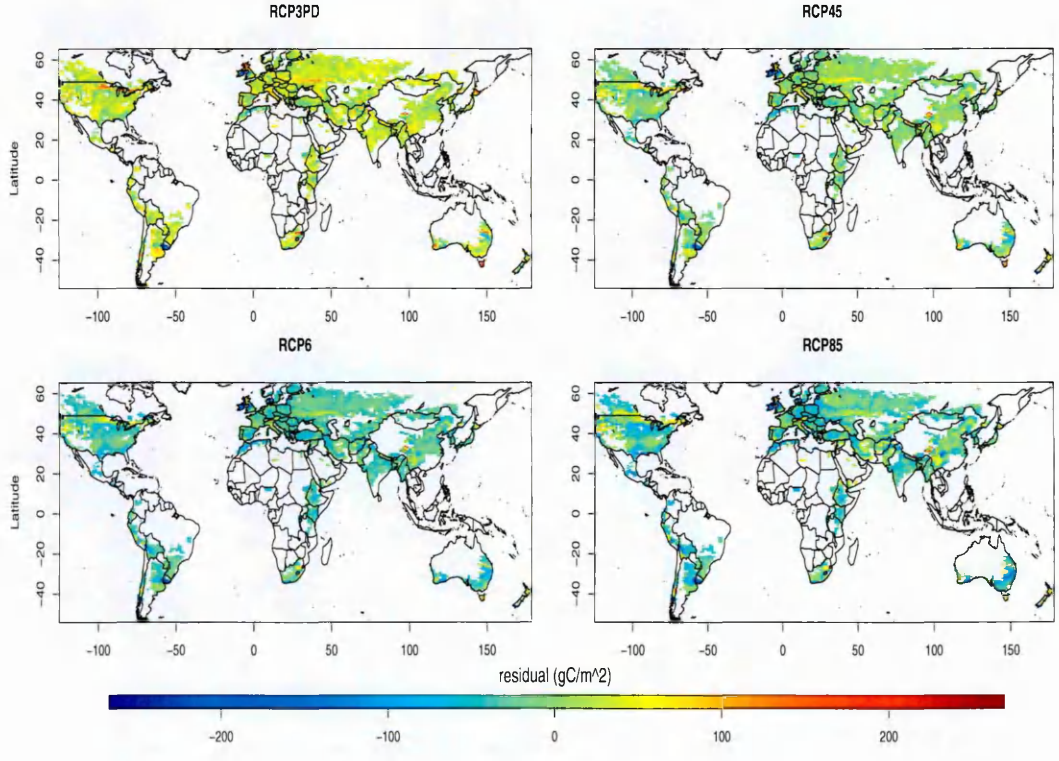


(b) Without CO_2 fertilization effect

Figure 7.6: Residual map for cereal change between (2085-2094) and (2005-2014), management level 5 with and without CO_2 fertilization effect from CCSR-MIROC32HI



(a) With CO_2 fertilization effect



(b) Without CO_2 fertilization effect

Figure 7.7: Residual map for cereal change between (2085-2094) and (2005-2014), management level 5 with and without CO_2 fertilization effect from CCCMA-CGCM31

Second stage summary

(i) Perform a principal components analysis of $\tilde{\mathbf{Y}}^{\mathrm{T}}\tilde{\mathbf{Y}}$. The non-zero eigenvalues are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{16}$ and the corresponding eigenvectors are $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{16}$. Put $\tilde{\mathbf{x}}^\star = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{16})^{\mathrm{T}}\tilde{\mathbf{y}}^\star$ and $\tilde{\mathbf{x}}_i = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{16})^{\mathrm{T}}\tilde{\mathbf{y}}_i$ for $i = 1, \ldots, 16$.

(ii) Denote the j^{th} components of $\tilde{\mathbf{x}}^\star$ and $\tilde{\mathbf{x}}_i$ by $x_j^\star$ and x_{ij}, respectively. Then $w_1, \ldots, w_{16}$ are the non-zero elements of the diagonal matrix $\mathbf{W}$, where $w_i = (1/d_i^2)/\{\sum_1^{16}(1/d_i^2)\}$, with $d_i^2 = \sum_{j=1}^{16}\lambda_j(x_j^\star - x_{ij})^2$.

(iii) The explanatory variables for the WLS regression are constructed from the first four eigenvectors of $\boldsymbol{\Gamma}$. We put $\hat{\boldsymbol{\Gamma}} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, \boldsymbol{\gamma}_4)$ and $\mathbf{X}_0 = \tilde{\mathbf{Y}}^{\mathrm{T}}\hat{\boldsymbol{\Gamma}}$.

(iv) Weighted least squares gives $\hat{\boldsymbol{\beta}}_n = (\mathbf{X}_0^{\mathrm{T}}\mathbf{W}\mathbf{X}_0)^{-1}\mathbf{X}_0^{\mathrm{T}}\mathbf{W}\tilde{\boldsymbol{\varepsilon}}_n$ as the vector of regression coefficients for the n^{th} grid cell, where $\tilde{\boldsymbol{\varepsilon}}_n^{\mathrm{T}}$ is the n^{th} row of $\mathbf{E}$.

(v) The estimated error for the n^{th} grid cell is $\hat{\varepsilon}_n^\star = \hat{\boldsymbol{\beta}}_n^{\mathrm{T}}\mathbf{x}_0^\star$, where $\mathbf{x}_0^\star = \hat{\boldsymbol{\Gamma}}^{\mathrm{T}}\tilde{\mathbf{y}}^\star$.

Put

$$\hat{\varepsilon}_n^{\#} = \begin{cases} \varepsilon_n^{min} & \text{if } \hat{\varepsilon}_n^\star < \varepsilon_n^{min} \\ \varepsilon_n^\star & \text{if } \varepsilon_n^{min} \leq \hat{\varepsilon}_n^\star \leq \varepsilon_n^{max} \\ \varepsilon_n^{max} & \text{if } \hat{\varepsilon}_n^\star > \varepsilon_n^{max}. \end{cases} \tag{7.3}$$

A Gaussian process model could not be applied directly to our data because of the computational difficulty from the large sample size coupled with the large number of parameters to be estimated. GP scales cubically with the number of observations $O(N^3)$, which is not appropriate for our present data – even after averaging decadally and sampling from each scenario, the data matrix contains approximately 4.5 million values. It might be possible to use GP for residual interpolation, rather than WLS, but this would still have a high computational cost and it would be necessary to reduce the resolution and aggregate data to a country level in order to reduce the computational load. However, we shall demonstrate how we can use GP regression to handle these data in Chapter 8.

7.3 Results

7.3.1 Cross-validation results

Figure 7.8 gives density plots over grid cells for the percentage change in crop yield (2084-2095) relative to the baseline yield for both CCSR-MIROC32HI and CCCMA-CGCM31 GCMs. We consider all management levels and a moderate emission scenario, RCP 6. The right-hand plots show the change in crop yield when there is no CO_2 fertilization while the left-hand plots show change in yield with CO_2 fertilization. In the right-hand plots, most crops show a preponderance of negative values, indicating a general reduction in yield. The exception is oil, which mainly shows positive values. The distributions each have a single major peak but vary as to whether they have a further minor peak (or even several minor peaks). The skewness of the density function also varies markedly with crop.

Now considering the left-hand plots (with CO_2 fertilization), the density plots show a preponderance of positive values and are more diverse in pattern, which could be a result of non-linear interactions between climate and CO_2. Comparison of the left- and right-hand plots shows that CO_2 fertilization has a marked effect on all crops except maize, for which the two plots are strikingly similar. (Maize is less affected by CO_2 fertilization because it is a C_4 plant and has a mechanism to efficiently transport CO_2 to the photosynthetic parts, limiting photorespiration rate thereby reducing water losses). Overall, the varying patterns in Figure 7.8 clearly show the diversity of the effects we are emulating. The changes in crop yields are characterized by high variability and there are varying patterns across different scenarios.

Figure 7.9 gives time series plots showing the percentage change in global crop yield over each decade relative to the baseline period under the CCSR-MIROC32HI, either without CO_2 fertilization (right-hand plots) or with CO_2 fertilization (left-hand plots). It shows temporal variability of the rainfed crops

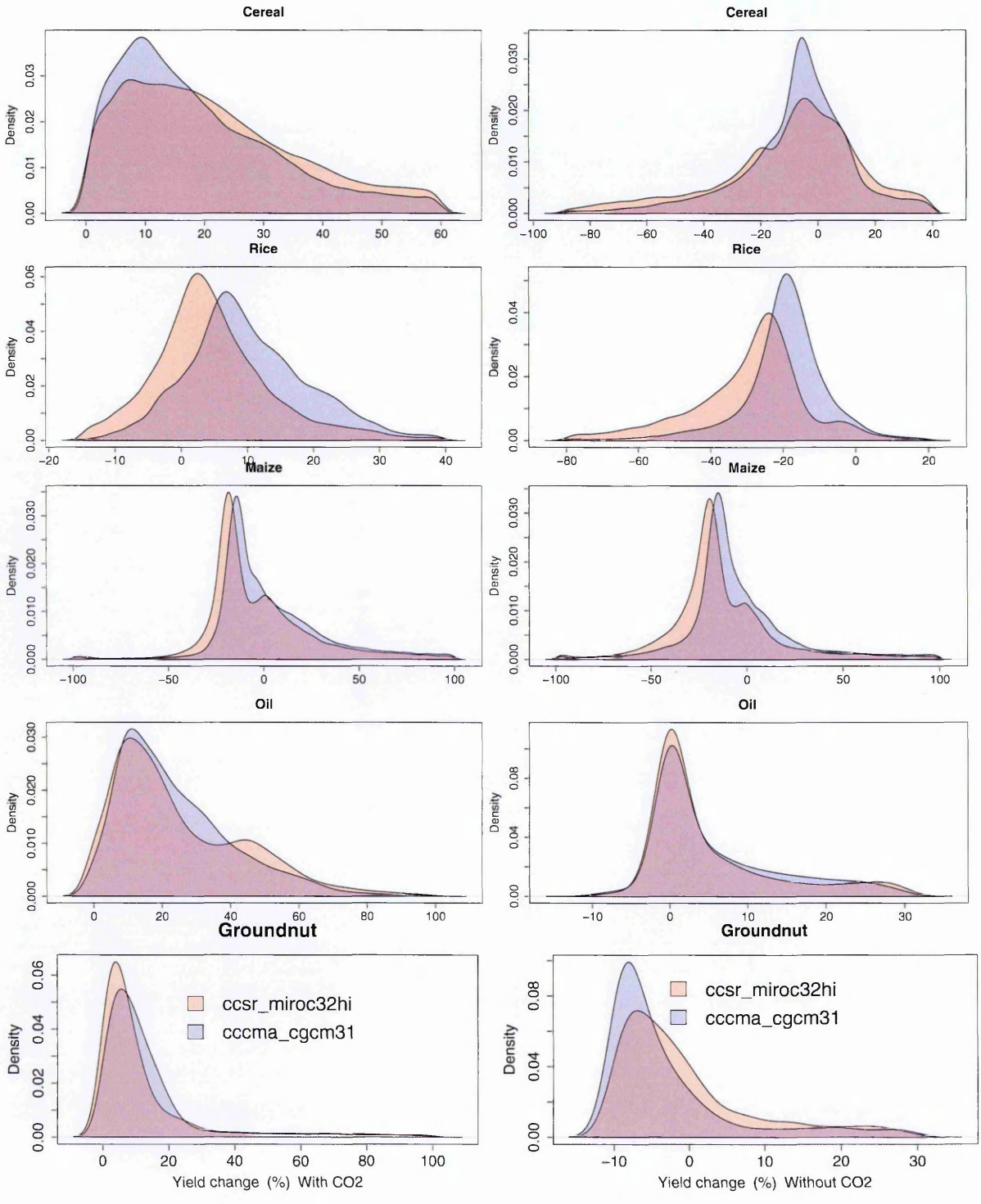


Figure 7.8: Probability distribution for the percentage decadal change between (2085-2094) and (2005-2014) for rainfed cereal, rice, maize, groundnut and oil respectively, RCP 6 and all management levels. Left-hand plots: with CO_2 fertilization; right-hand plots: without CO_2 fertilization.

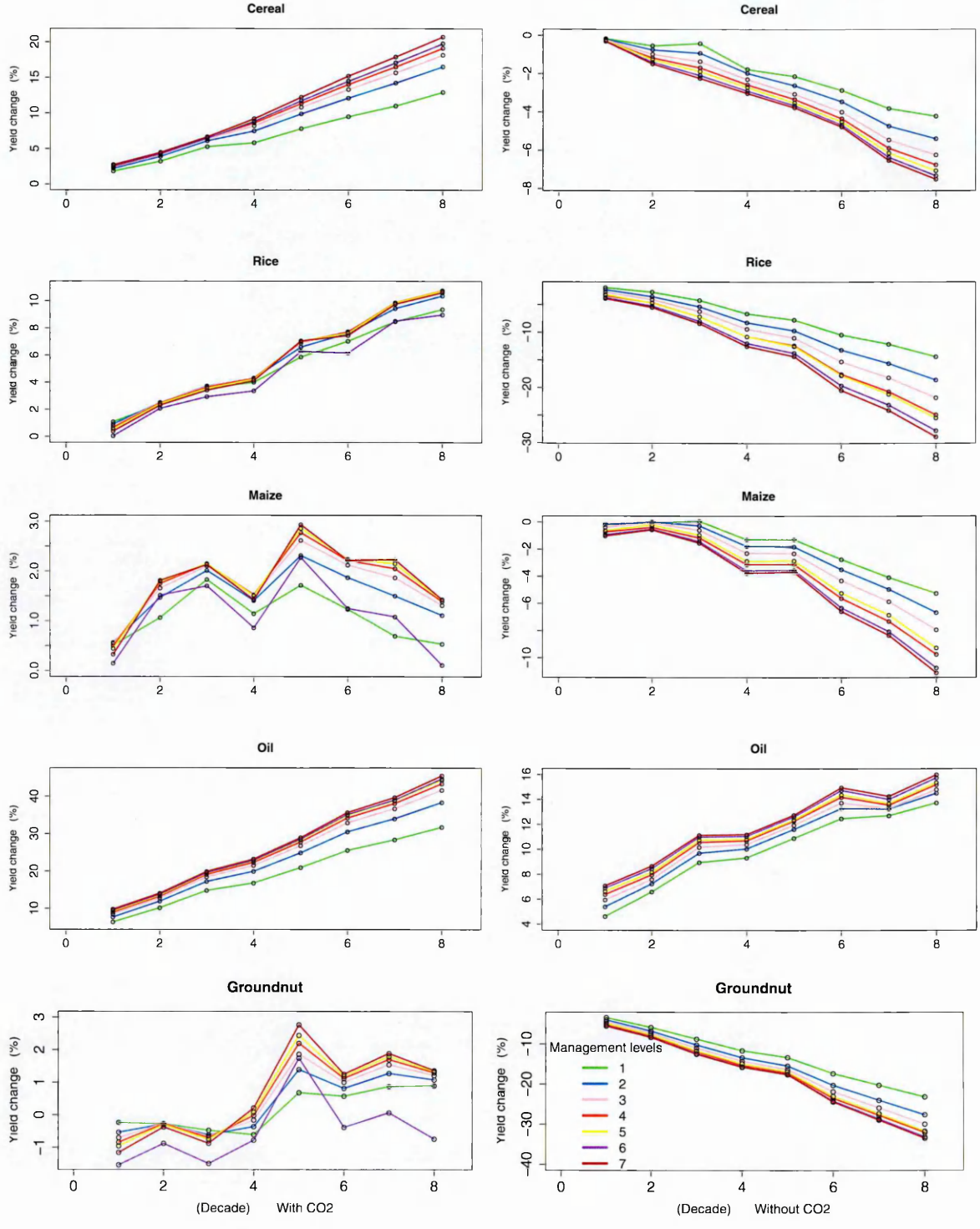


Figure 7.9: Time series plot showing the temporal pattern for the percentage decadal change relative to baseline period for rainfed temperate cereal, rice, maize, groundnut and oil respectively, for all time-slices, RCP 6 and all management levels for the CCSR-MIROC32HI average over all grid cells. Left-hand plots: with CO_2 fertilization; right-hand plots: without CO_2 fertilization.

Table 7.3: Cross-validated proportion of variance ρ and root mean squared error RMSE_{CV} showing the overall performance of the emulators for rainfed and irrigated crops, with all management levels, RCPs, and time slices, but with CO_2 fertilization only, for UKMO-HADGEM1.

Crop	1^{st} stage ρ		2^{nd} stage ρ		RMSE_{CV} (gC/m^2)	
	rainfed	irrigated	rainfed	irrigated	rainfed	irrigated
Cereal	0.41	0.41	0.60	0.64	16.72	14.94
Rice	0.39	0.37	0.62	0.78	14.71	24.02
Maize	0.35	0.45	0.74	0.80	17.79	20.45
Oil[1]	0.51	0.49	0.73	0.81	12.34	7.65
Groundnut	0.24	0.40	0.62	0.79	12.04	17.72

[1] Oil=yield_{max}[soybean, rapeseed, sunflower].

under the RCP6 scenario with a separate line for each management level. The sensitivity of change in yield to the CO_2 fertilization effect is apparent, with CO_2 fertilization greatly improving the change in yield of most crops. Taking management level 4 and the last decade (2085-2094) as an example, for cereal a decline in yield of 6% becomes a growth of 18%; rice and oil show improvements of 37% and 25%, respectively, while groundnut has an increase of 32.5%. Maize exhibit a weak sensitivity to CO_2 , with the globally averaged yield increasing by 11%. When there is no CO_2 fertilization, with all management levels there is a fairly steady reduction in yield for cereal, rice, maize and groundnut while oil shows an increase in yield. When there is CO_2 fertilization, maize and groundnut yield still change comparatively little over the decades and the effect of random variation is more apparent in their time series plots. Oil, unlike the other crop categories, increases with or without CO_2 fertilization but at a much lower rate when CO_2 fertilization is absent.

Table 7.3 summarises the cross-validated performance of the emulators for stage 1 and stage 2 using equation 5.72, for both rainfed and irrigated crops under UKMO-HADGEM1. Maize and oil cross-validated noticeably better than other crops with 74%/73% variance explained when rainfed and 80%/81% explained

Table 7.4: Cross-validated proportion of variance ρ for four GCMs, with CO_2 fertilization, management level 5, RCPs 4.5 and 8.5, and all time slices

Crop	CCSR-MIROCMED		GISS-MODELER		GISS-MODELEH		IPSL-CM4	
	rainfed	irrigated	rainfed	irrigated	rainfed	irrigated	rainfed	irrigated
Cereal	0.79	0.82	0.67	0.62	0.72	0.80	0.74	0.83
Rice	0.79	0.91	0.75	0.77	0.68	0.85	0.75	0.88
Maize	0.86	0.85	0.72	0.66	0.75	0.76	0.83	0.87
Oil[1]	0.88	0.93	0.69	0.78	0.72	0.84	0.81	0.82
Ground	0.78	0.89	0.70	0.69	0.70	0.82	0.71	0.84

[1] Oil=yield$_{max}$[soybean, rapeseed, sunflower].

when irrigated.

Generally, the emulator of the irrigated crops performed better than the emulator of the rainfed crops. This could be attributed to the water stress in rainfed locations, difficult to model, that could complicate the predictions. Stage 1 explained less than 50% of the variance except for rainfed oil. However, the second stage of the algorithm improved results for both the rainfed and irrigated crop systems. For the rainfed crops, the values of the variance explained increased from (24-51%) to (60-74%) for all the crops. For the irrigated crops, the first stage explained variance was between 37-49% for all the crops and this increased to 64-81%, as shown in Table 7.3.

The last two columns of Table 7.3 show the computed RMSE$_{CV}$ for all scenarios, time slices and management levels, which further examines the accuracy of the emulators. RMSE$_{CV}$ is the difference between the LPJmL and emulator predictions and provides a measure of uncertainty associated with the emulator. Irrigated oil and rice have the lowest and highest values with 7.65 and 24.02 gC/m^2 respectively; a low value indicates more accurate predictions.

The cross-validation of stage 2 predictions for four additional GCMs are shown in Table 7.4. We can see that the emulators again performed well, with re-

sults that are typically a little better than in Table 7.3. The results for CCSR-MIROCMED are better than for other GCMs; this was to be expected because a very similar GCM, CCSR-MIROC32HI, gave part of the training data. Results for irrigated crops are better than for rainfed crops, which was also the case in Table 7.3, though the results with GISS-MODELER for cereal, maize and groundnut are an exception.

We now consider the spatial comparison between LPJmL and the emulators of UKMO-HADGEM1 for temperate cereal. The map in Figure 7.10 displays results for both rainfed and irrigated temperate cereals. The emulator under-predicts yield change across the United States for rainfed cereal and over-predicts yield of the irrigated crop in some regions, especially in Eastern Asia and Europe. Overall, the emulator reproduces the global patterns well, especially for the irrigated crop ($\rho = 0.69$ for rainfed cereal and 0.74 for irrigated cereal).

Figure 7.11 shows results for the rice emulator. The emulator for the rainfed crops under-predicts the rice almost everywhere except in Eastern Asia where it reproduces the pattern quite well. On the other hand, an emulator for the irrigated crops reproduces the yield better than the rainfed ($\rho = 0.74$, 0.90 respectively). There is a potential for higher irrigated rice yield in Europe and Asia as shown by both the LPJmL and irrigated emulator plots. Higher yield changes are more prominent with irrigated rice than for rainfed. More irrigated rice is grown than rainfed rice especially in latitude $\geq 30°$, this area is also characterized by a high change in yield.

Similar to what we observed for cereal and rice, the rainfed crops emulator under-predicts maize in Europe and some part of Russia (7.12). The emulator relatively well predicts all other areas. The irrigated crop emulator captures the spatial patterns of maize very well. Higher yield changes are associated with rainfed maize than for irrigated especially at latitudes $\geq 50°$. Figure 7.13 shows oil plots, the rainfed emulator under-predicts oil in some part of Russia while other regions are well predicted by the emulator. The rainfed LPJmL pattern

is quite different to the irrigated LPJmL pattern. Oil values are over-predicted in most areas across the globe by the irrigated emulator. Groundnut results are shown in Figure 7.14, the emulator seems to over-predict LPJmL values in Africa and some part of Asia. Other regions are quite well predicted.

Overall, rainfed crop patterns are quite different from the irrigated, as expected, because irrigation allows some crops to be grown where they would not have grown naturally (e.g. rice is grown in Europe with irrigation). Also, more negative changes are prominent in both the LPJmL and emulator predictions for rainfed rice than for other crops. We can clearly see that the emulators cross-validated well as indicated by their ρ values (Tables 7.3 and 7.4) and thus captured relatively well the spatial patterns of LPJmL. The maps show visually that the emulator produces patterns that are quite similar to the LPJmL patterns, although there are over- and under-predictions in some instances.

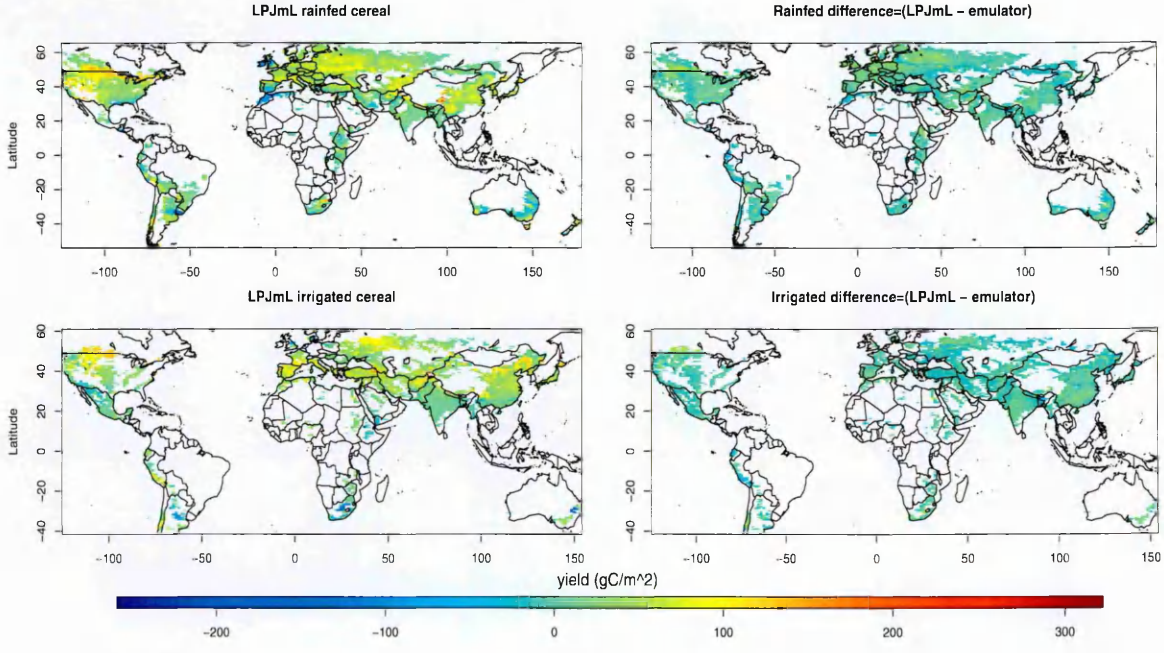


Figure 7.10: Cross-validation of the comparison between LPJmL and difference between LPJmL and predictions for rainfed and irrigated temperate cereals, plotted as mean decadal change in yield between (2085-2094) and (2005-2014). The plots correspond to crop-yield for management level of 5, RCP6 with CO_2 fertilization from UKMO-HADGEM1. The white over land correspond to grid cells with zero observations.

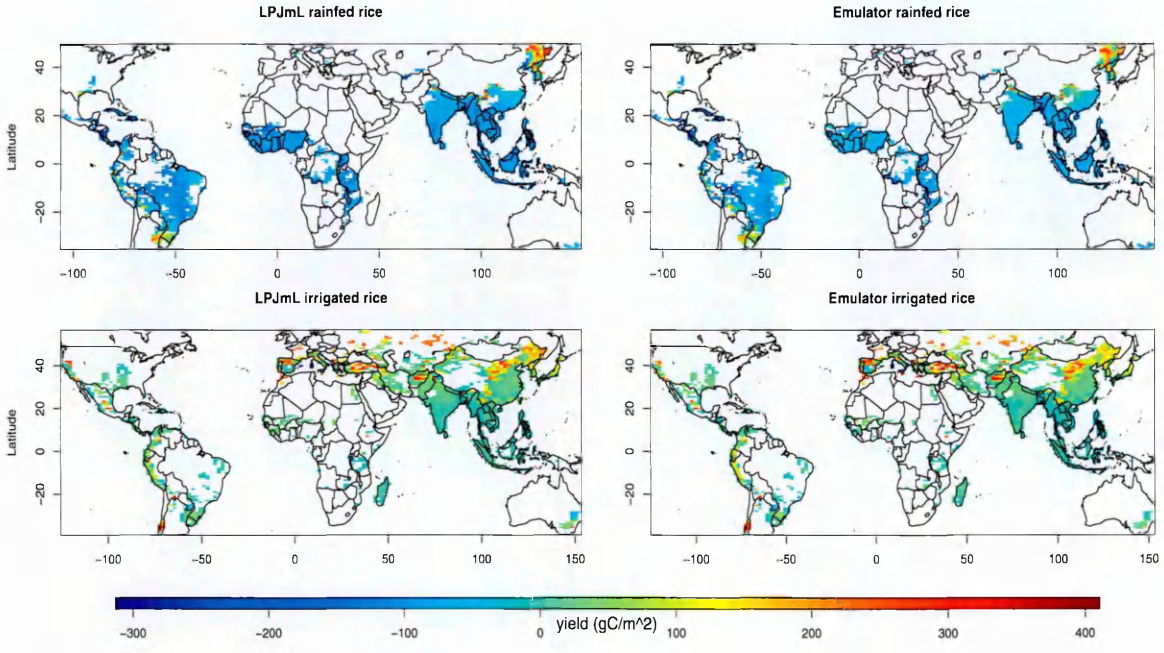


Figure 7.11: Cross-validation of the comparison between LPJmL and LPJmLem for rainfed and irrigated rice. This plot is mean decadal change in yield between (2085-2094) and (2005-2014). The plots correspond to crop-yield for management level of 5, RCP6 with CO_2 fertilization from UKMO-HADGEM1. The white over land correspond to grid cells with zero observations.

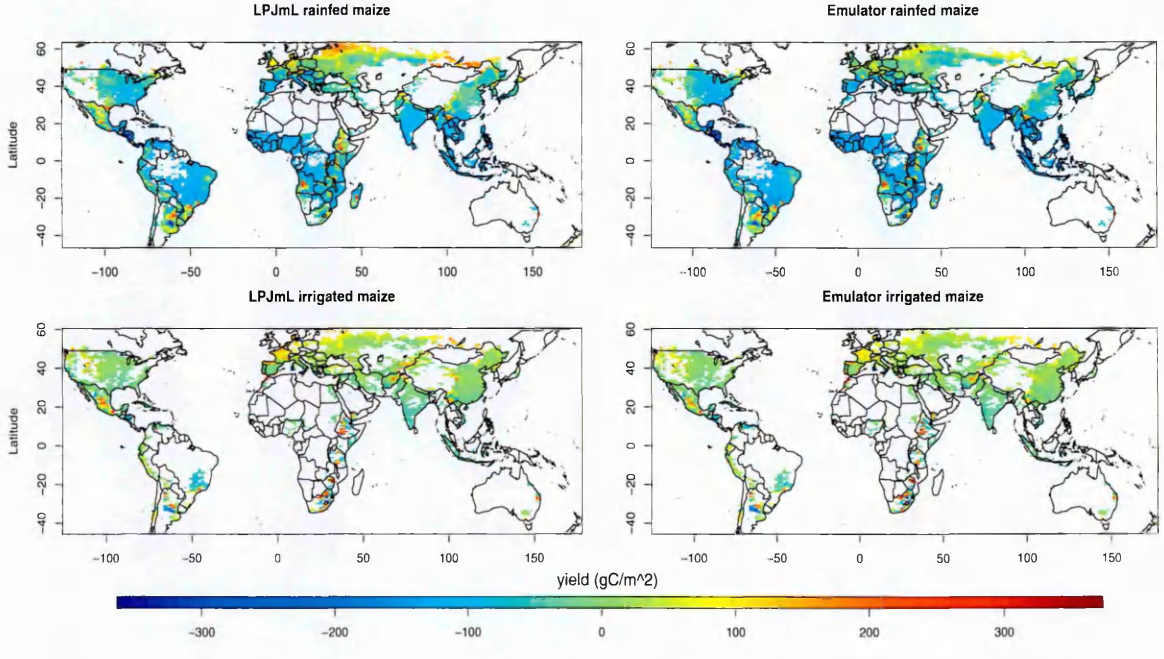


Figure 7.12: Cross-validation of the comparison between LPJmL and LPJmLem for rainfed and irrigated maize. This plot is mean decadal change in yield between (2085-2094) and (2005-2014). The plots correspond to crop-yield for management level of 5, RCP6 with CO_2 fertilization from UKMO-HADGEM1. The white over land correspond to grid cells with zero observations.

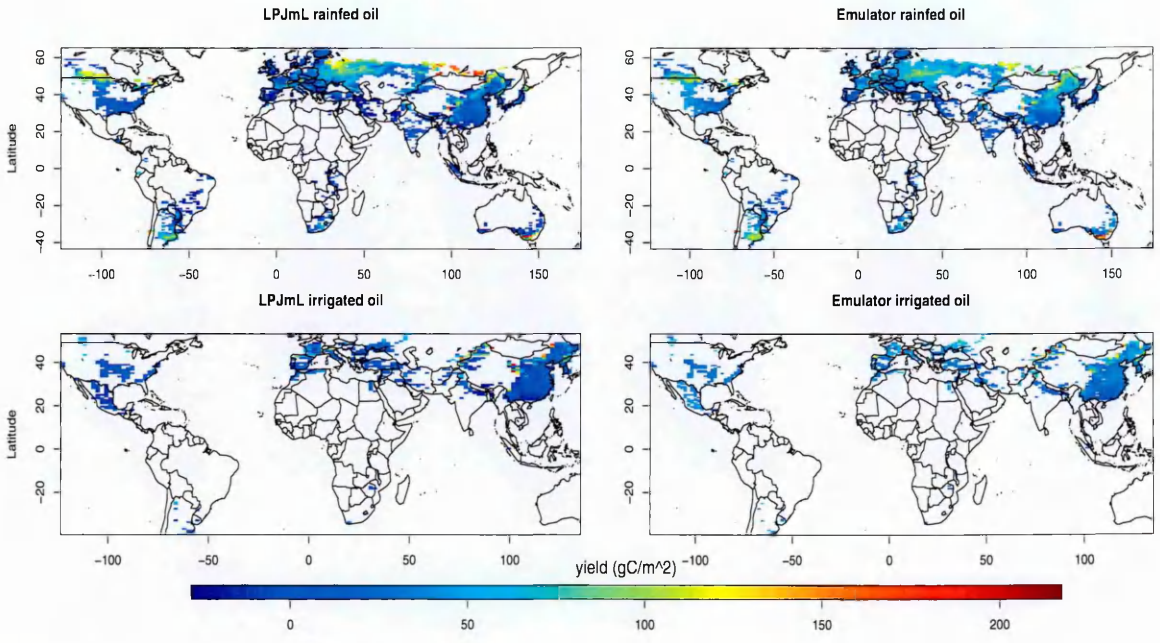


Figure 7.13: Cross-validation of the comparison between LPJmL and LPJmLem for rainfed and irrigated oil. This plot is mean decadal change in yield between (2085-2094) and (2005-2014). The plots correspond to crop-yield for management level of 5, RCP6 with CO_2 fertilization from UKMO-HADGEM1. The white over land correspond to grid cells with zero observations.

7.3.2 Variograms and distance metrics

Distance weighted regression is used to estimate the residual pattern of an unknown scenario from the scenarios with known residual patterns (Section 7.2.3). More weight is assigned to known scenarios that are similar in pattern to the unknown scenario. The distance, d_i, between the unknown ith (known) scenario is defined in equation (6.5) and is taken as the measure of pattern similarity.

Three simple metrics for converting the d_i distances to weights were considered: linear ($w_i \propto d_i^{-1}$), quadratic ($w_i \propto d_i^{-2}$), and cubic ($w_i \propto d_i^{-3}$), where the weights (w_i) are scaled so that $\sum w_i = 1$. The use of covariance functions to determine weights was also explored using variograms. For each crop/irrigation regime/time slice/ management level there are sixteen different training scenarios that predict the yield in a grid cell. Generalising equation (6.5), let $d_{ik} = \left[\sum_{j=1}^{16} \lambda_j (x^{\star}_{ij} - x_{kj})^2\right]^{1/2}$ denote the distance between the ith and kth scenarios, and let z_{ikn} denote the difference between their prediction errors in grid

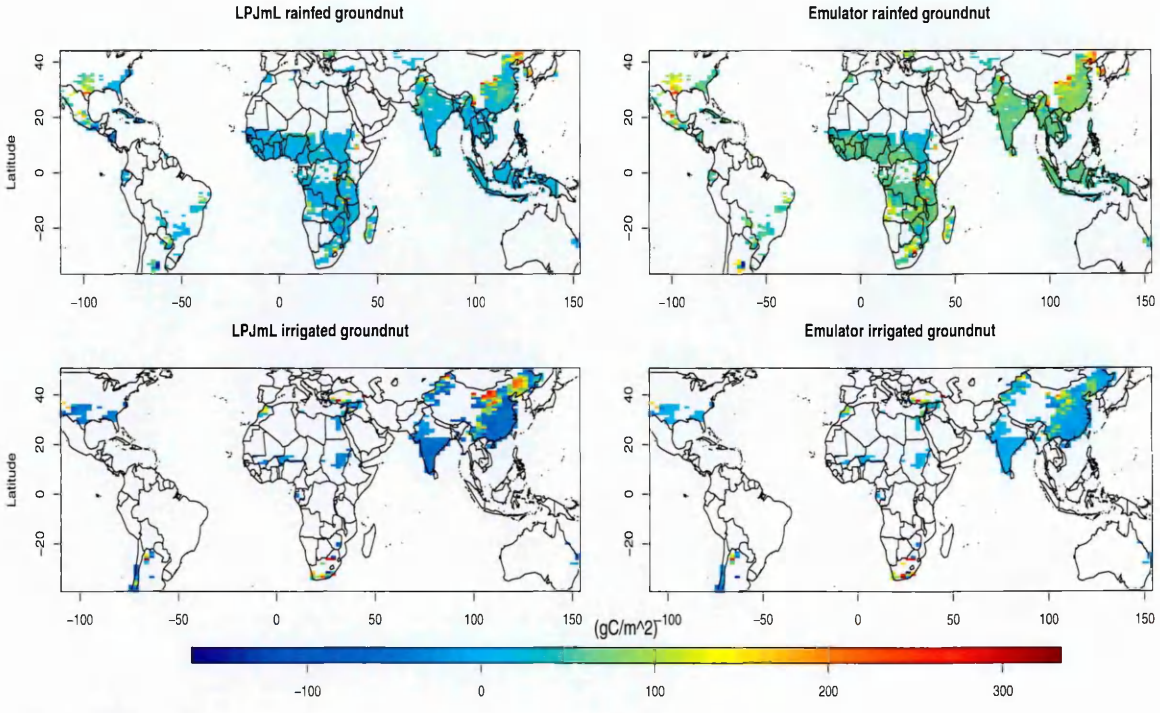


Figure 7.14: Cross-validation of the comparison between LPJmL and LPJmLem for rainfed and irrigated groundnut. This plot is mean decadal change in yield between (2085-2094) and (2005-2014). The plots correspond to crop-yield for management level of 5, RCP6 with CO_2 fertilization from UKMO-HADGEM1. The white over land correspond to grid cells with zero observations.

cell n.

To obtain the empirical variogram, we split the range of d_{ik} into sections (bins) of equal length. The empirical variogram, $\hat{\gamma}(d)$ is defined as (Cressie, 1993)

$$\hat{\gamma}(\overline{d}_\ell) = \frac{1}{M_\ell N} \sum_{(i,k)\in\Omega_\ell} \sum_{n=1}^{N} |z_{ikn}|$$

where $\overline{d}_\ell$ is the middle of bin ℓ, $(i,k) \in \Omega_\ell$ if d_{ik} is in bin ℓ, and M_ℓ is the number of items in Ω_ℓ. In Figure 7.15 the small circles are the values of $\hat{\gamma}(\overline{d}_\ell)$ at the midpoint of each bin. The lines in the figure show the theoretical variograms for the following covariance models.

The parameters σ and ϕ are estimated from the parametric variograms given in Chapter 5 section 5.6.5. For the Matern model, κ must be specified and we consider three values, $\kappa = 0.1$, $\kappa = 0.5$ and $\kappa = 2$.

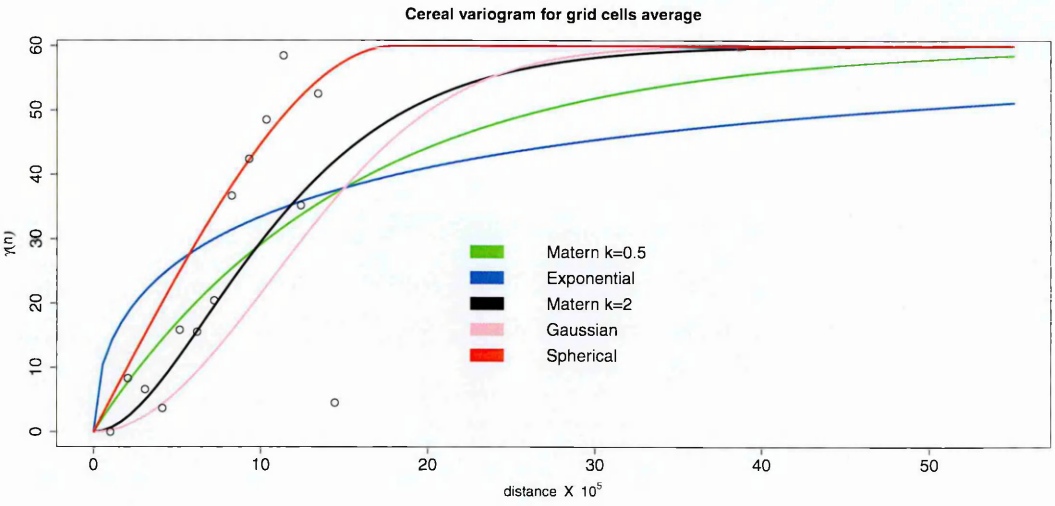


Figure 7.15: Empirical and theoretical variogram for rainfed cereal. The points are the estimated variogram bins using the residual data, while the curves are the theoretical models fitted using various covariance models.

Figure 7.15 shows the variogram plot for rainfed cereal. Here, the residuals are averaged over all the grid cells. The plotted points in Figure 7.15 show that $\hat{\gamma}(\overline{d}_\ell)$ increases rapidly as $\overline{d}_\ell$ increases from 0 (and then levels off), so it is clear that errors for the different scenarios are correlated and the correlation increases as distance reduces. There are variations in ability of each parametric model to capture all the points. Variogram plots for different crops are given in the Appendix (Figures B.1). These show the same trend. For all crops, the theoretical covariance models follow the empirical points reasonably closely with no model being clearly the best. Additional variogram plots for some randomly selected grid cells are also shown in Appendix Figure B.2 that again show an increase in $\hat{\boldsymbol{\gamma}}(\overline{d}_\ell)$ as $\overline{d}_\ell$ increases.

A covariance model assigns weights to scenarios so that $w_i \propto \{\gamma(d_i)\}^{-1}$ where, as with the other metrics, weights are scaled so that $\sum w_i = 1$. The emulator was fitted using different weight functions, enabling the covariance models and other metrics to be further compared using cross-validation. Results are presented in Table 7.5. The exponential covariance model led to poor predictions for rainfed oil and rainfed groundnut, but otherwise all the methods of choosing weights led to reasonably good predictions. The quadratic distance metric gave somewhat

Table 7.5: Cross-validated proportion of variance ρ for various covariance functions used as a weight in WLS fitting compared to quadratic metric with management level 5, RCP 6 between (2085-2094) and (2005-2014).

Function	Cereal		Rice		Maize		Oil[1]		Groundnut	
	rain[3]	irrig[2]	rain	irrig	rain	irrig	rain	irrig	rain	irrig
Quadratic	0.69	0.74	0.74	0.90	0.81	0.86	0.80	0.88	0.68	0.79
Spherical	0.76	0.82	0.78	0.84	0.83	0.73	0.75	0.84	0.68	0.79
Matern $\kappa = 0.5$	0.78	0.82	0.78	0.84	0.84	0.87	0.79	0.88	0.67	0.79
Matern $\kappa = 2$	0.78	0.82	0.80	0.85	0.85	0.88	0.79	0.88	0.68	0.81
Matern $\kappa = 0.1$	0.79	0.82	0.78	0.83	0.83	0.88	0.79	0.88	0.67	0.79
Exponential	0.76	0.81	0.65	0.89	0.79	0.86	0.45	0.82	0.59	0.85
Gaussian	0.76	0.80	0.76	0.68	0.77	0.56	0.70	0.79	0.66	0.75

[1] Oil=yield$_{max}$[soyabean, rapeseed, sunflower], [2] "irrig" denotes irrigated crop
[3] "rain" denotes rainfed crop.

better results than the covariance models in Table 7.5 and it explained 68-90% of the variation in the LPJmL predictions, compared with 45-89% for the covariance models. Because of its simplicity and comparable performance, we chose it as the means of determining weights for the WLS regression in stage 2 of the emulator algorithm. We observe that Matern results changes with respect to the choice of smoothing parameter κ except for irrigated cereal, and rainfed and irrigated oil.

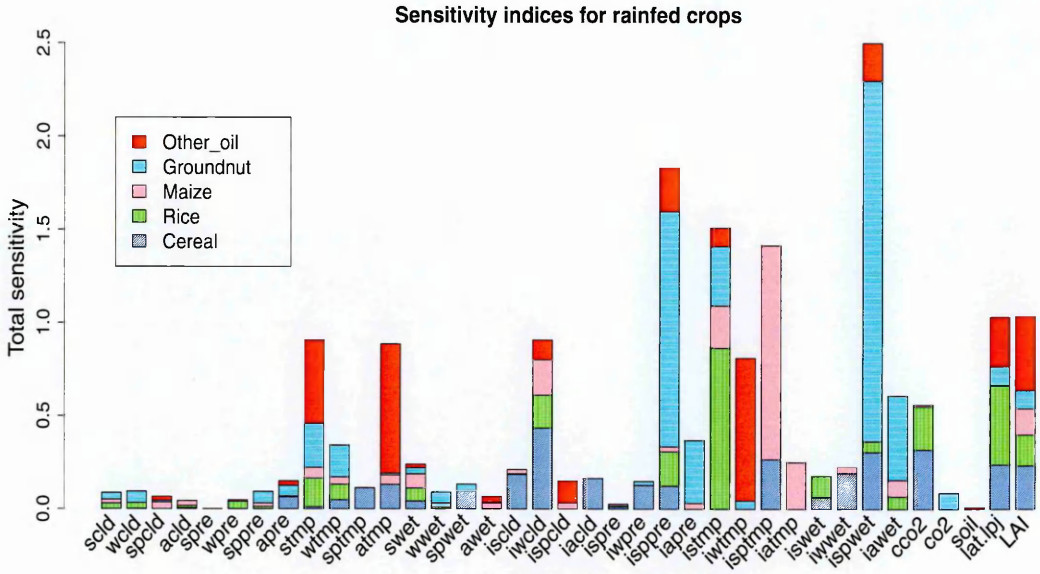
7.3.3 WLS diagnostic results

Diagnostics plots for WLS regression are shown in Appendix Figures C.1 and C.2. The random errors are the unexplained variation left after applying the WLS to the residual from stage 1 (i.e observed residual minus expected residual). The regression assumptions hold reasonably well. Figure C.1a (top-left) plots residual against the fitted values and is relatively good as there are no obvious patterns, with the points randomly scattered. QQ plot (top-right), the extreme values deviate from the straight line at both ends, but the sample-size is only 16 and the Q-Q plot is close to normal away from the extremes. The plot of a

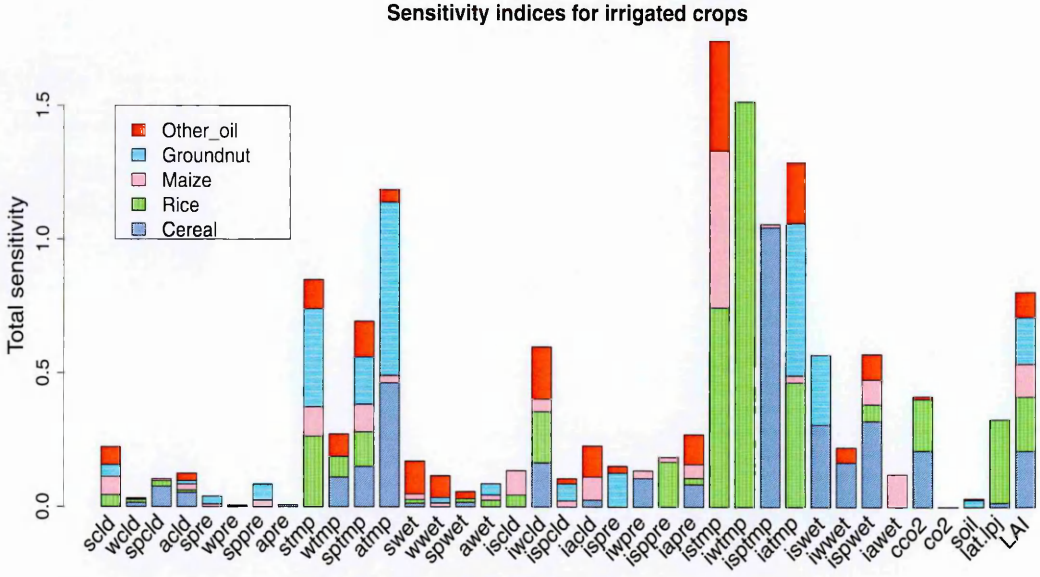
standardized residual against the fitted values (bottom-left) gives no indication of variance heterogeneity, with no increasing or decreasing patterns. In general, there is no obvious problem with these plots. Similar results are observed for equivalent plots from other randomly chosen grid points Figures C.1b, C.2a and C.2b.

7.4 Sensitivity results

Here, we investigate how the uncertainty in the crop-yield can be partitioned to the various uncertainties in the input variables. We sampled 20 000 observations directly from the simulation with the CO_2 effect for each of the 37 input variables. We computed both the first order (results are not shown) and total sensitivity indices as described briefly in Chapter 5 section 5.8. The Bootstrapping technique was used to compute the 95% Confidence Interval on the estimated indices. This procedure was applied to all the five crops for both rainfed and irrigated crops. The "sensitivity" package in R (2013) was used for this analysis. Total sensitivity results are shown in Figure 7.16a for rainfed crops. Results for irrigated crops are given in Figure 7.16b



(a) Rainfed



(b) Irrigated

Figure 7.16: Barplots showing the sensitivity indices for the five rainfed crops over all time-slices, RCPs and GCMs (negative indices are set to 0). See Table D.1 for full names.

The six most relevant parameters for prediction of temperate cereal yields are initial winter cloud cover, CO_2 change, initial spring wetday frequency, initial spring temperature ('initial' represents baseline value), latitude and LAI, with indices of 0.43, 0.32, 0.30, 0.27, 0.23 and 0.23, respectively. The parameter with the second highest rank is CO_2 change, reflecting the sensitivity of yield to the CO_2 fertilization effect. Atmospheric CO_2 improves crop productivity by stimulating

photosynthesis, thus increasing the number of fruits, seeds. The fourth most sensitive parameter is initial spring temperature, which generally has a considerable effect on plant growth.

Initial summer temperature, latitude, CO_2 change, LAI, initial spring precipitation are the five most important variables for rice. Their indices are 0.86, 0.42, 0.23, 0.18 and 0.18, making it very clear that summer temperature is the most important parameter for rice. It is twice as influential as latitude (although temperature and latitude are obviously correlated) and four times more influential than CO_2 change. Rice is a C_3 plant that utilizes direct carbon fixation of CO_2, so CO_2 change is expected to be an important parameter.

The most relevant variables for maize are initial spring, autumn and summer temperature, as well as initial winter cloud cover and LAI. Their indices are 1.1, 0.25, 0.22, 0.19 and 0.14, respectively. It is unsurprising that, as these results show, seasonal temperatures play a key role in the growth and development of maize plants. Maize is less affected by CO_2 fertilization because it is a C_4 plant which has a more efficient mechanism to transport CO_2 to photosynthetic paths. Oil has the following important variables: initial winter temperature, autumn and summer temperature change, LAI and latitude with indices of 0.85, 0.78, 0.40, 0.26 and 0.20, respectively.

Overall, we can see from Figure 7.16a that baseline seasonal temperature, spring precipitation and wetday frequency, LAI, latitude, CO_2 change and cloud cover are the most important variables across the five crops. Although CO_2 change has a negligible effect on maize, groundnut and oil, our results further support the joint interactive effect of elevated CO_2 and temperature on crop-yield. Some variables are less important (examples are the change in seasonal precipitation and cloud cover), and some are unimportant (such as summer precipitation and soil) because their calculated sensitivity indices are very low. Baseline CO_2 is of limited importance in this analysis because its level is represented by a single global value and does not vary with time-slice, RCP, or GCM. Similarly, soil is

represented in this study by discrete values that range from 1-8 across the globe and these are constant across RCP and GCM scenarios.

Figure 7.16b shows the results of a sensitivity analysis for irrigated crops. The most sensitive parameters for cereal are initial spring temperature, change in autumn temperature, initial spring wetday, initial summer wet day, CO_2 change and LAI, with indices of 1.05, 0.46, 0.32, 0.31, 0.21 and 0.21, respectively. Similarly, rice is heavily dependent on temperature as indicated by the indices for initial winter temperature, initial summer temperature, initial autumn temperature and change in summer temperature with indices of 1.52, 0.75, 0.47 and 0.26. But it is less sensitive than cereal to latitude, LAI and CO_2 change, with indices of 0.31, 0.20 and 0.19.

Maize is determined primarily by initial summer temperature (0.59) and less dependent on precipitation and cloud cover, but it is non-sensitive to CO_2 change. Similarly, oil is sensitive to seasonal temperature, like other crops, but it is not dependent on CO_2 and CO_2 change. Irrigated crops are less dependent on precipitation than rainfed crops but rely more on temperature. Also, irrigated crops are less sensitive to latitude, soil, and CO_2 as compared to rainfed in Figure 7.16a.

In general, these sensitivity analyses result clearly indicated that temperature is the most uncertain parameter in crop yield projection. This high sensitivity of (growing season) temperature was earlier observed as the major determinant of crop yield change under future climate by Lobell & Burke (2008) and Osborne et al. (2013).

7.5 Conclusion

The objective of this chapter was to provide a means of emulating crop yield patterns under climate change. We have shown that this can be done using OLS-PCA-WLS combinations, including parametric covariance functions that introduce much flexibility to our distance metrics. It was demonstrated that emulation can provide reliable and useful predictions for the LPJmL.

In comparison to other crop modelling studies, our emulation approach extends the work of Lobell & Burke (2010, 2008) with incorporation of other covariates like soil, latitude, crop management levels in addition to climate variables for crop predictions. Specifically, our approach integrated the three methods of time series, panel and cross-sectional analyses described in Lobell & Burke (2010) for crop predictions. We performed extensive cross-validation for five climate models and four RCPs unlike Lobell & Burke (2010) that do not test the validity of their statistical models in other sites not included in the training data.

The global sensitivity analysis in this chapter provides a measure of the contribution of each variable to the overall uncertainty. Apart from using the popular variance based decomposition technique, the results can be additionally used for future calibration especially for LPJmL with numerous parameters. In accordance with Lobell & Burke (2008), our sensitivity results also indicated that temperature is a greater source of uncertainty than other variables for future crop impacts assessment.

Chapter 8

Bayesian emulation of crop yields

8.1 Introduction

Bayesian methods, particularly Gaussian process (GP), have been prominent in the construction of emulators. Our simulation data for both crops and climate are high-dimensional gridded global data that cannot be modelled directly by GP. Several approximation techniques have been developed to handle multi-dimensional data. For instance, a low-rank approximation of the Gram matrix was described in Rasmussen & Williams (2006); Drineas & Mahoney (2005a). The use of subsets of regressors and subsets of data as an iterative solution under a linear model, to reduce the computational burden of the GP inversion for a large data matrix was investigated by (Silverman, 1985; Wahba, 1990; Wahba et al., 1995). Several studies have used Bayesian techniques for emulating high-dimensional data (Rougier, 2008; Rougier et al., 2009; Heitmann et al., 2006; Higdon et al., 2008; Dancik, 2011). In particular, Bornn & Zidek (2012); Finley et al. (2011) used Bayesian methods for predicting crops.

Bayesian methods typically offer wide applicability and flexibility. In the context of emulation, a further benefit is that they can help quantify model uncertainty. A disadvantage is that, computationally, they can be very demanding. This is a serious drawback for the emulation problems addressed in this thesis for the following reasons.

1. Climate varies considerably from one spatial scale to another. This variability is determined largely by atmospheric circulation and its interactions with large-scale ocean currents. Regional or local climate is much more variable than global climate because climate on a large spatial scale is less influenced by internal dynamics of the continental or global climate. Therefore, even if we decide to use a PC decomposition or aggregation of the crop yield data (to reduce the dimension), we would not be able to apply that data reduction strategy to climate input parameters as they vary from one location to another. The decomposition of the data is non-trivial as climate may vary a lot within a country so that the average climate across a country may be a poor approximation to the local climates within the country. In addition, LPJmL data is a non-linear response so large scale average climate may be an inappropriate input to our problem.

2. The LPJmL data that we are using are monthly data and they have been averaged to decadal level. We have already lost some information due to decadal averaging because the LPJmL model itself works on a daily time step. We do not want to loose further vital information from the data by further averaging or by applying principal components to reduce the dimension of our data. Applying PCA to reduce dimension of LPJmL will result in PC components that ignore information about the climate input variables thereby resulting in PC components that may be difficult to explain by the original input variables.

These drawbacks limit the ways in which we can simplify our problem so that GP method can be incorporated in an emulator. In Chapter 7, a two-stage technique was proposed for emulating LPJmL crop data. In the first stage of that method, OLS regression was applied onto crop yield, and WLS was used for residual interpolation in the second stage. A Gaussian process model could not be applied directly to the first stage because of the computational difficulty in the sample size coupled with the large number of parameters to be estimated. GP

scales cubically with the number of observations $O(N^3)$, which is not appropriate for our present data, even after averaging decadally and sampling from each scenario. The data matrix contains approximately 4.5 million values. Following Lee et al. (2013) and Lee et al. (2012), we considered treating each spatial point as a GP and then emulating each cell individually. However, this option is intractable since we have 59199 grid points meaning that about 236,796 emulators would have to be constructed and validated for the four crops we are emulating.

It is possible to replace WLS with GP in the second stage of the emulator and that is the focus of this chapter. However, the approach still has a high computational cost, and it is necessary to reduce the spatial resolution and aggregate data to a country level in order to reduce the computational load.

We will compare the PCA-WLS method that we described in Chapter 7 with a PCA-GP model. In particular, we demonstrate the use of a PCA-GP model in emulating the crop yield response under global climate scenarios. We examine whether the non-parametric modelling of the unexplained residual using Gaussian process (GP) in our second stage provides a more accurate result than the weighted linear regression described in the last chapter. In this thesis, we focus only on emulator uncertainty rather than parametric uncertainty of the input space.

8.2 GP emulation procedure

Our objective is to produce a statistical approximation that will link the model climate inputs $\mathbf{X}$ to the deterministic vector outputs $\mathbf{y} = f(\mathbf{X})$ from LPJmL. Recall from Chapter 7, the relationship between the climate input data from MAGICC6/ClimGen and each output from LPJmL is represented by a model

$$\mathbf{y} = f(\mathbf{X}) + \varepsilon,$$

where $\mathbf{y}$ is the (vector) simulated mean decadal change for LPJmL crop-yield over the grid cells for all scenarios, with $f(\mathbf{X})$ as defined by equation (6.1) in Chapter 7. The vector $\boldsymbol{\varepsilon}$ represents all the unexplained variation in $f(\mathbf{X})$ which is not captured by the OLS method in the stage 1 emulator. Values of $\boldsymbol{\varepsilon}$ are spatially correlated for scenarios that are very close in the input space. In this section, we model $\boldsymbol{\varepsilon}$ as a Gaussian random process with known mean and covariance functions. GP is an extension of multivariate Gaussian distributions to infinite dimensionality with associated mean and covariance matrix.

Consider a single crop/irrigation regime/management level/time slice combination and let $\mathbf{y}_i$ be the vector of changes in yield given by LPJmL for that combination in the i^{th} scenario ($i = 1, \ldots, 16$). As in Chapter 7, we let $\tilde{\mathbf{y}}_i$ be the corresponding predictions given by the stage 1 emulator and $\boldsymbol{\varepsilon}_i = \mathbf{y}_i - \tilde{\mathbf{y}}_i$ is the error in prediction. Each $\tilde{\mathbf{y}}_i$ and $\boldsymbol{\varepsilon}_i$ is an $N \times 1$ vector, where N denotes the number of grid cells for that crop/irrigation regime. As $\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{16}$ are predictions on *training* data, the values of $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_{16}$ are known.

Given a new vector of predictions, $\tilde{\mathbf{y}}^{new}$, from a new scenario where the emulator values are unknown, the aim is to estimate the error of $\tilde{\mathbf{y}}^{\star}$ from $\mathbf{E} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_{16})$ and $\tilde{\mathbf{Y}} = (\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{16})$. In order to reduce the dimension of data we apply a PC decomposition to the $16 \times N$ matrix $\tilde{\mathbf{Y}}^{\mathrm{T}}$, and select just the first four principal components. The resulting 16×4 matrix of PC scores given by these four components is denoted as $\tilde{\mathbf{U}}$. The columns of matrix $\tilde{\mathbf{U}}$ will be used as explanatory variables for the GP regression of residual patterns. In Chapter 7, WLS was used instead of GP regression.

Let $\boldsymbol{\varepsilon}^{new}$ denote the error in $\tilde{\mathbf{y}}^{new}$ after stage 1. A separate GP regression is used for each component of $\boldsymbol{\varepsilon}^{new}$. The dimension of the residual matrix $\mathbf{E}^{\mathrm{T}}$ is $16 \times N$, ($N = 16648$ for rainfed cereal) which is very large for the purposes of GP regression. As each GP regression is computationally intensive, performing many GP regressions is impractical. However, often the results from an emulator are only required at the level of individual countries, rather than at the level

of individual grid cell. As there are only 186 separate countries (excluding very small countries), performing a separate GP regression for each country is feasible and that is what we do in this chapter. To this end, we apply spatial aggregation to reduce the dimension to a country level using equation (8.10). The countries we consider are listed in Table 8.1.

The spatial aggregation will transform $\mathbf{E}^{\mathrm{T}}$ to matrix $\mathbf{Z}$ of dimension 16×186 ($\mathbf{E}^{T}_{16\times N} \rightarrow \mathbf{Z}_{16\times 186}$). Details of the aggregation are provided under section 8.3. So for the n^{th} GP regression (the n^{th} country) the values of the dependent variable are $\mathbf{z}_n$, where $\mathbf{z}_n$ is the n^{th} column of $\mathbf{Z}$ (ie $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_{186})$) is a vector of observed residuals for the n^{th} country.

After aggregating the residual to a country level, we take one country at a time and form a separate GP regression equation for that country. The data for one of these regressions is the (16×1) vector of responses $\mathbf{z}$ (the residuals for that country) and the (16×4) matrix $\tilde{\mathbf{U}}$, which holds the values taken by the explanatory variables.

For the new scenario, put $\mathbf{u}^{new} = \hat{\mathbf{\Gamma}}^{\mathrm{T}}\tilde{\mathbf{y}}^{new}$. Then the estimate of the prediction error in $\tilde{\mathbf{y}}^{new}$ at the n^{th} country is derived as following. Set up the Gaussian process by selecting a mean and a covariance function. Then the residual $\mathbf{z}$ can be modelled as a GP regression given by

$$\mathbf{z} = g(\mathbf{u}) = \mathbf{h}^{\mathrm{T}}(\mathbf{u})\boldsymbol{\beta} + \delta(\mathbf{u}), \tag{8.1}$$

where $\mathbf{h}(\mathbf{u})$ is a vector of regression functions and is chosen to reflect the functional form of our response data. In a GP regression, $\mathbf{h}(.)$ can be either a constant mean function ($\mathbf{h}(.) = 1$, corresponding to only an intercept term) or it can be modelled as a simple linear regression function of the inputs ($\mathbf{h}(\mathbf{u}) = (1, \mathbf{u}^T)$, including an intercept term). We used $\mathbf{h}(\mathbf{u}) = (1, \mathbf{u}^T$ in this thesis.

The vector $\boldsymbol{\beta}$ is an unknown hyperparameter to be estimated and $\delta(.)$ is a stationary GP representing stochastic noise with mean zero and covariance

Table 8.1: List of UN countries used in the analyses of Chapter 8

Index	Country	Index	Country	Index	Country
1	Ocean	63	French Polynesia	125	New Zealand
2	Afghanistan	64	French Southern Territories	126	Nicaragua
3	Albania	65	Djibouti	127	Niger
4	Antarctica	66	Gabon	128	Nigeria
5	Algeria	67	Georgia	129	Norway
6	Angola	68	Gambia	130	Pakistan
7	Azerbaijan	69	State of Palestine	131	Panama
8	Argentina	70	Germany	132	Papua New Guinea
9	Australia	71	Ghana	133	Paraguay
10	Austria	72	Kiribati	134	Peru
11	Bahamas	73	Greece	135	Philippines
12	Bangladesh	74	Greenland	136	Poland
13	Armenia	75	Guatemala	137	Portugal
14	Belgium	76	Guinea	138	Guinea-Bissau
15	Bhutan	77	Guyana	139	Timor-Leste
16	Bolivia	78	Haiti	140	Puerto Rico
17	Bosnia and Herzegovina	79	Honduras	141	Qatar
18	Botswana	80	Hungary	142	Romania
19	Brazil	81	Iceland	143	Russian Federation
20	Belize	82	India	144	Rwanda
21	Solomon Islands	83	Indonesia	145	Saint Vincent and Grenadines
22	Brunei Darussalam	84	Iran	146	Saudi Arabia
23	Bulgaria	85	Iraq	147	Senegal
24	Myanmar	86	Ireland	148	Serbia
25	Burundi	87	Israel	149	Sierra Leone
26	Belarus	88	Italy	150	Slovakia
27	Cambodia	89	Cote d'Ivoire	151	Viet Nam
28	Cameroon	90	Jamaica	152	Slovenia
29	Canada	91	Japan	153	Somalia
30	Cape Verde	92	Kazakhstan	154	South Africa
31	Cayman Islands	93	Jordan	155	Zimbabwe
32	Central African Republic	94	Kenya	156	Spain
33	Sri Lanka	95	Democratic Republic of Korea	157	South Sudan
34	Chad	96	Republic of Korea	158	Sudan
35	Chile	97	Kuwait	159	Suriname
36	China	98	Kyrgyzstan	160	Swaziland
37	Taiwan	99	Lao Republic	161	Sweden
38	Colombia	100	Lebanon	162	Switzerland
39	Comoros	101	Lesotho	163	Syrian Arab Republic
40	Congo	102	Latvia	164	Tajikistan
41	D.R Congo	103	Liberia	165	Thailand
42	Costa Rica	104	Libya	166	Togo
43	Croatia	105	Lithuania	167	Trinidad and Tobago
44	Cuba	106	Luxembourg	168	United Arab Emirates
45	Cyprus	107	Madagascar	169	Tunisia
46	Czech Republic	108	Malawi	170	Turkey
47	Benin	109	Malaysia	171	Turkmenistan
48	Denmark	110	Mali	172	Uganda
49	Dominican Republic	111	Mauritania	173	Ukraine
50	Ecuador	112	Mauritius	174	The former Yugoslav
51	El Salvador	113	Mexico	175	Egypt
52	Equatorial Guinea	114	Mongolia	176	United Kingdom
53	Ethiopia	115	Republic of Moldova	177	United Republic of Tanzania
54	Eritrea	116	Montenegro	178	United States of America
55	Estonia	117	Morocco	179	United States Virgin Islands
56	Faeroe Islands	118	Mozambique	180	Burkina Faso
57	Falkland Islands (Malvinas)	119	Oman	181	Uruguay
58	South Georgia and South SS	120	Namibia	182	Uzbekistan
59	Fiji	121	Nepal	183	Venezuela
60	Finland	122	Netherlands	184	Samoa
61	Aland Islands	123	New Caledonia	185	Yemen
62	France	124	Vanuatu	186	Zambia

function

$$K = Cov(\mathbf{z}(\mathbf{u}), \mathbf{z}(\mathbf{u}^{\mathrm{T}})) = \sigma^2 C(\mathbf{u}, \mathbf{u}^{\mathrm{T}}),$$

where $C(\mathbf{u}, \mathbf{u}^{\mathrm{T}})$ is a correlation function. The covariance function K must be a semi-positive definite function to ensure that it can be inverted. We chose a squared exponential correlation function that decreases as the distance between $\mathbf{u}$ and $\mathbf{u}^{\mathrm{T}}$ increases. It has the form

$$\mathbf{C} = Cor(\mathbf{u}, \mathbf{u}^{\mathrm{T}}) = \exp^{\left[\sum_{j=1}^{4} -\boldsymbol{\alpha}_j(\mathbf{u}_j - \mathbf{u}_j^{\mathrm{T}})^2\right]}. \tag{8.2}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_4)$ is a smoothness parameter that measures the rate of change in the response as the input value changes; it will be estimated from the training data.

The Gaussian process has the same functional form as the multivariate Gaussian distribution. If we assume a Gaussian prior process for the output function $g(.)$ and update this with our training data $\mathbf{D}^{\mathrm{T}} = [(z_i, g(u_{ij})); i = 1, \ldots, 16; j = 1, \ldots, 4]$, we obtain a posterior distribution that is also a Gaussian distribution. GP regression will be used to estimate z^{new}, the corresponding error of $\mathbf{u}^{new}$. In order to predict the response at a new input point $\mathbf{u}^{new}$, given some training data $\mathbf{D}$, the joint distribution of the observed values $\mathbf{D}$ and test point z^{new} can be obtained using a multivariate Gaussian identity. Suppose that

$$z^{new}|\mathbf{D} = \sim N\left[\begin{pmatrix} h^T \\ \mathbf{H} \end{pmatrix}\boldsymbol{\beta}, \begin{pmatrix} \mathbf{c(u)} & \mathbf{t(u^{\mathrm{T}})} \\ \mathbf{t(u)} & \mathbf{C} \end{pmatrix}\sigma^2\right] \tag{8.3}$$

where $\mathbf{t(u)} = Cor(\mathbf{u}, \mathbf{u}^{new})^{\mathrm{T}}$ is a correlation vector of training data $\tilde{\mathbf{U}}$ with a new input point $\mathbf{u}^{new}$, $\mathbf{C}$ is a 16×16 correlation matrix among the training data $\tilde{\mathbf{U}}$, $\mathbf{c} = Cor(\mathbf{u}^{new}, (\mathbf{u}^{new})^{\mathrm{T}})$ is the correlation between the test data. Then, the conditional posterior distribution has the form

$$P(z^{new}|\mathbf{D}, \boldsymbol{\beta}, \sigma^2 \sim N(\mu^{\bullet}, \mathbf{K}^{\bullet}) \tag{8.4}$$

and after some algebraic manipulation (by integrating out the hyperparameters)

the posterior mean and covariance functions are given respectively as

$$\mu^{\bullet}(\mathbf{u}) = h^T(\mathbf{u})\widehat{\boldsymbol{\beta}} + \mathbf{t}(\mathbf{u})\mathbf{C}^{-1}[\mathbf{D} - \mathbf{H}(\mathbf{u})\widehat{\boldsymbol{\beta}}] \tag{8.5}$$

$$K^{\bullet}(\mathbf{u}, \mathbf{u}^{\mathrm{T}}) = \hat{\sigma}^2\Big(\mathbf{c}(\mathbf{u}) - \mathbf{t}(\mathbf{u}^T)\mathbf{C}^{-1}\mathbf{t}(\mathbf{u})\Big). \tag{8.6}$$

These equations directly follow from subsection 5.6.3 in Chapter 5, where $\mathbf{H}$ is the 16×4 matrix given as $\mathbf{H}^{\mathrm{T}} = [\mathbf{h}(\mathbf{u}_1), \ldots, \mathbf{h}(\mathbf{u}_{16})]$, with $\mathbf{u}_i$ denoting the i^{th} design point and hyperparameters $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\alpha}$ are obtained using maximum likelihood. The mean function $\mu^{\bullet}(\mathbf{u})$ is an estimate of the function $g(\mathbf{u})$ and the covariance function $K^{\bullet}(\mathbf{u}, \mathbf{u}^{\mathrm{T}})$ is a measure of uncertainty associated with this mean function.

$$\hat{\boldsymbol{\beta}} = (\mathbf{H}\mathbf{C}^{-1}\mathbf{H})^{\mathrm{T}}\mathbf{H}\mathbf{C}^{-1}\mathbf{z} \tag{8.7}$$

$$\widehat{\sigma^2} = \frac{1}{n}\Big[(\mathbf{z} - \mathbf{H}\hat{\boldsymbol{\beta}})^T\mathbf{C}^{-1}(\mathbf{z} - \mathbf{H}\hat{\boldsymbol{\beta}})\Big]. \tag{8.8}$$

Computation of the inverse and determinant of the correlation matrix $\mathbf{C}_{\mathbf{n}\times\mathbf{n}}$ is always tedious and time-consuming for large n which is one of the limitations of GP regression. Our analysis is fitted with the package *mlegp* documented in Dancik (2011). Further information on GP regression is given in subsection 5.6.3 of Chapter 5.

Having obtained both the posterior mean and estimated variance of the actual residual at a new input point for each country, we need to compute the final predictions from the emulator. These are then given as $\hat{\mathbf{y}}_{agg}^{new}$:

$$\hat{\mathbf{y}}_{agg}^{new} = \tilde{\mathbf{y}}_{agg}^{new} + \mu^{\bullet}(\mathbf{u}), \tag{8.9}$$

where $\tilde{\mathbf{y}}_{agg}^{new}$ is obtained by aggregating half degree data $\tilde{\mathbf{y}}^{new}$ to obtain values at country level. To combine the calculations for all countries into a single step, we fitted an independent GP regression in parallel to the columns of $\mathbf{Z}$.

We applied the same concept to model each of the four crops (cereal, rice, maize and oil). Groundnut is not emulated in this chapter. This process was re-

peated for all the time slices, RCPs and management levels. The training GCMs are CCSR-MIROC32HI and CCCMA-CGCM31. In order to test the performance of the emulator, a complete cross-validation was performed using another five selected GCMs, UKMO-HADGEM1, GISS-MODELER, GISS-MODELEH, IPSL-CM4 and CCSR-MIROCMED. Some cross-validation results are provided in section 8.4.

8.3 Spatial aggregation procedure

In order to apply the GP regression for residual interpolation, we reduce the resolution of the residuals data from half degree to a country level by aggregating the residual matrix $\mathbf{E}^{\mathrm{T}}$. Country-level residual values are computed by multiplying the residual at each grid cell by the crop growing area for each cell (separately for rainfed and irrigated crops). We calculated the area-weighted sum for each country to determine the total residual for that country, and finally divided this total residual by the total (country) area.

Therefore, for the i^{th} residual scenario,

$$\mathbf{Z}_{[i,n]} = \frac{\sum_{j=1}^{m} \mathbf{E}^{\mathrm{T}}_{[i,j]} \times \mathbf{A}_j}{\sum_{j=1}^{m} \mathbf{A}_j} \tag{8.10}$$

where m is the number of grid cells that fall into country n, $\mathbf{E}$ is the residual as defined earlier and $\mathbf{A}_j$ is the crop growing area for the j^{th} grid cell ($i = 1, \ldots, 16$; $j = 1, \ldots, N$). This analysis is performed with the *rworldmap* package in *R*. The package documents the aggregation of global half degree gridded data to a country level; for further detail see South (2011).

8.4 GP results

We choose one combination of crop/management/time point from a particular scenario to demonstrate the results. The results shown here are for the 8^{th} decade corresponding to change between the average yield in the period 2085-2094 and average yield in the baseline period average (2005-2014). The situation we consider is where the CO_2 fertilization effect is included. The climate at this time point is assumed to be characterized by a relatively large climate uncertainty associated with a high impact on yield change, and RCP 6 which is a moderate emission scenario under management level 5. Crop management level is a measure of vegetation density in cropping systems as influenced by machinery and fertilizer application. A well-managed system is assumed to have values of ≥ 5 for developed countries, while under-developed countries like Sub-Sahara Africa are assigned values between 1 and 3. Figure 8.1 shows the percentage of variance explained by each PC component for the four crops we consider in this chapter. We can see that the first 4 PCs explained at least 95% of the variance for each crop.

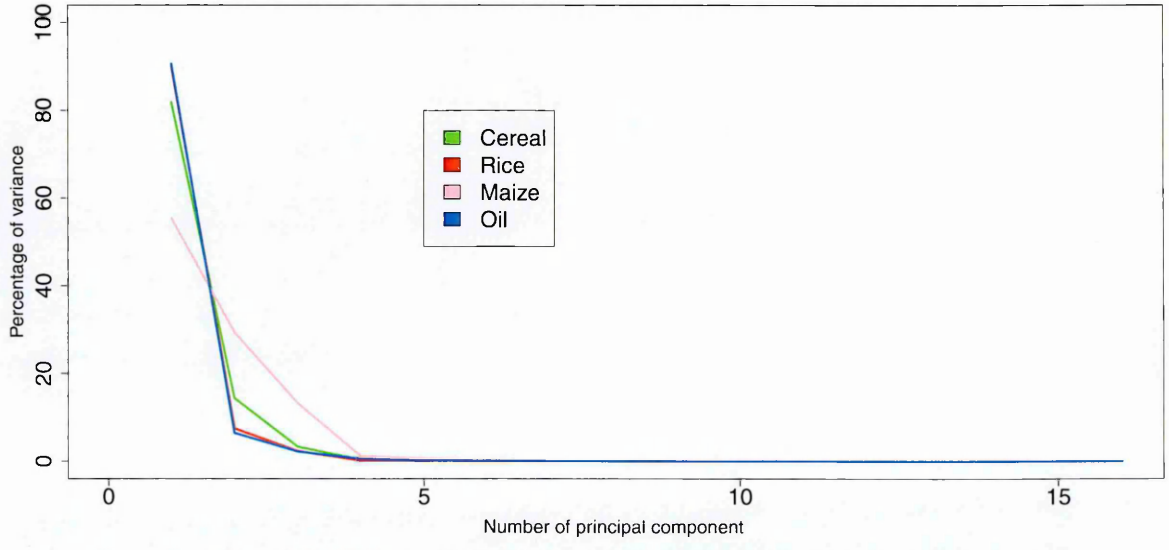
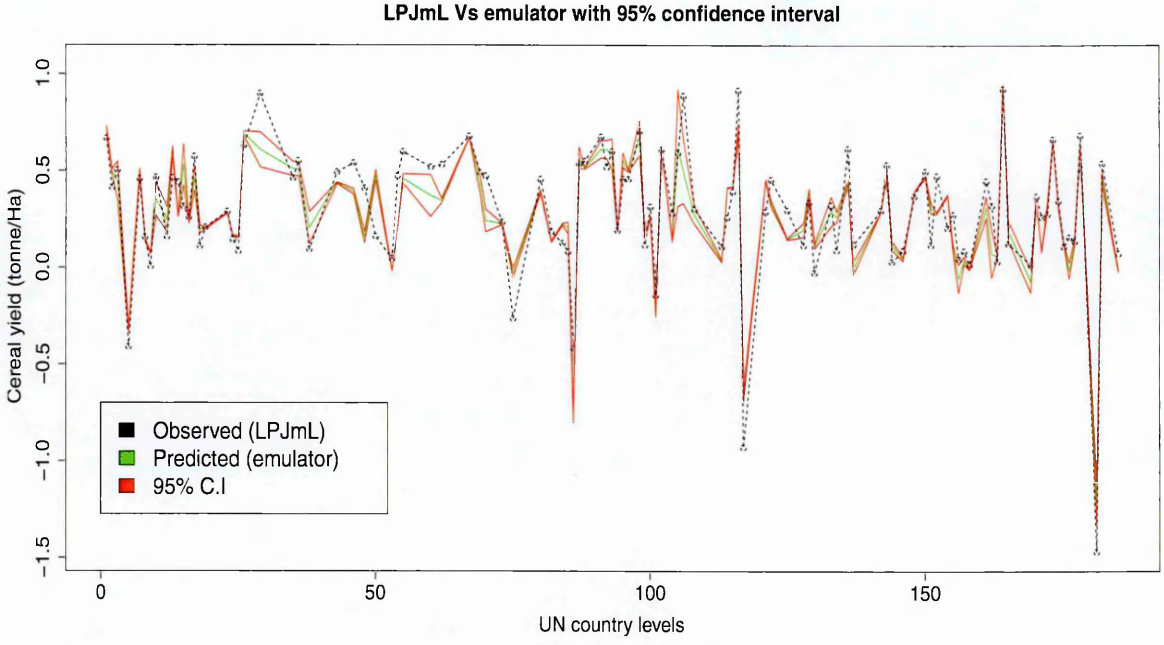
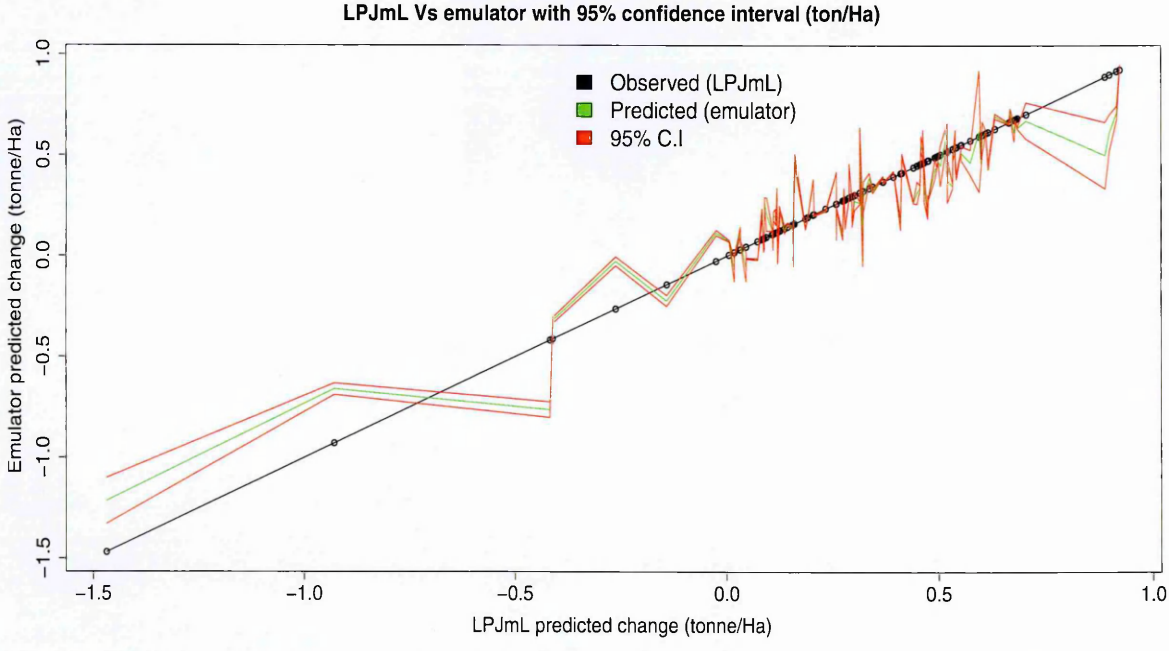


Figure 8.1: Percentage of the variance explained by each PC for the four rainfed crops. The plots correspond to crop-yield for management level of 5, RCP6 with CO_2 fertilization from UKMO-HADGEM1. [1] Oil=yield$_{max}$[soybean, rapeseed, sunflower].



(a) LPJmL and emulator predictions including their 95% C.I vs. UN country levels



(b) Pair plot of emulator predictions vs. LPJmL values including their 95% C.I

Figure 8.2: Cross-validation for rainfed cereal; LPJmL and emulator with its 95% C.I for each countries for mean decadal change in yield between (2085-2094) and (2005-2014) with their 95% C.I. The plots correspond to crop-yield for management level of 5, RCP6 with CO_2 fertilization from UKMO-HADGEM1.

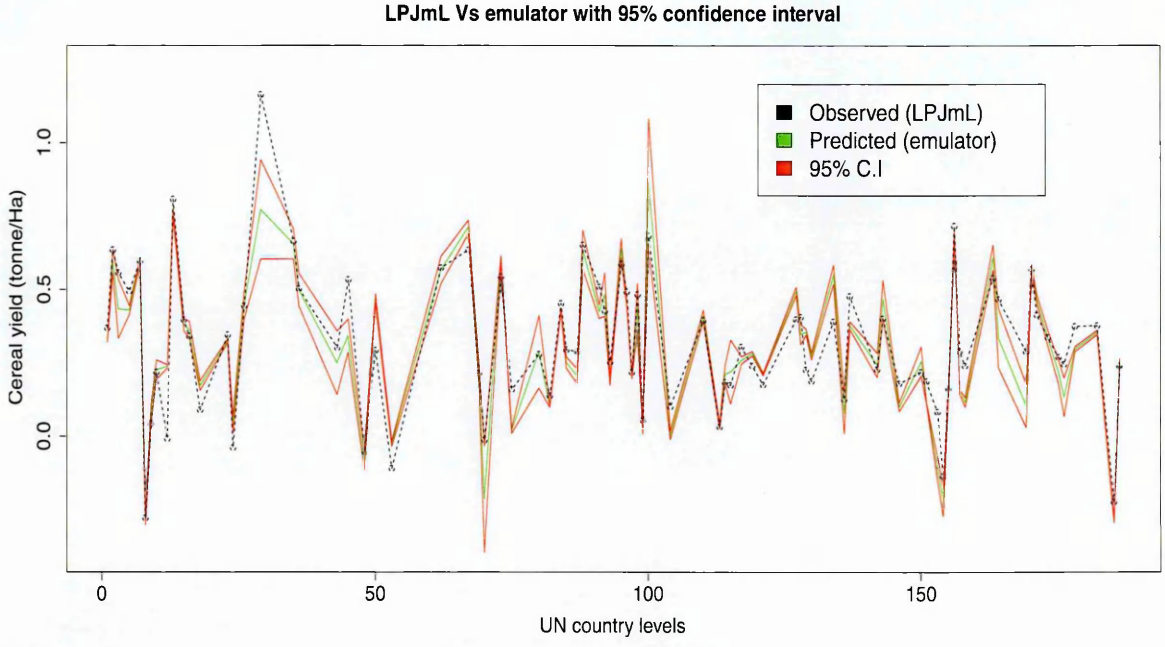
Figure 8.2 is the GP emulation results for the 8^{th} decadal change of rainfed temperate cereal under RCP6, management level 5 and 8^{th} decadal change for UKMO-HADGEM1. Figure 8.2a shows the performance of the 106 GP rainfed

emulators for the cereal. Each GP corresponds to an individual country where a rainfed cereal crop is grown.

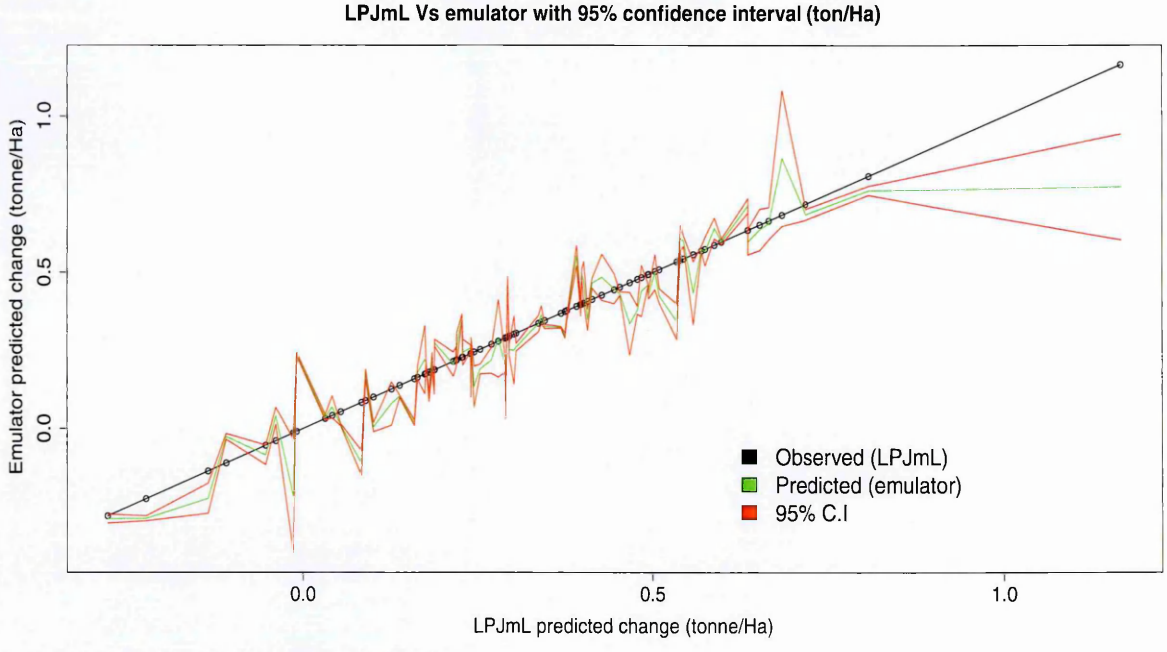
The four countries with the largest increase in yield are Tajikistan, Montenegro, Canada and Luxembourg. See Table (8.1) for a list of UN countries. Uruguay, Morocco, Ireland and Algeria are four countries with the greatest reduction in cereal yield (corresponding to the four largest negative peaks in Figure 8.2). There is relatively little increase in yield in other countries. The observed values for LPJmL are within the 95% C.I of the emulator predictions for 57 out of 106 countries considered which is equivalent to $\sim 54\%$. The emulator predicts Tajikistan relatively well, giving it the largest increase. Uruguay has the lowest change in yield and the change is under-predicted by the emulator; the confidence interval given by the emulator is -1.21 ± 0.05 tonne/Ha, while the value given by LPJmL is -1.46 tonne/Ha.

Figure 8.2b is a bilinear plot between LPJmL and the emulator for the 106 countries. The LPJmL values are arranged in ascending order from least to highest. The plot shows more clearly that the change in yield for LPJmL are clustered together for most of the countries. The two end points correspond to Tajikistan and Uruguay, and uncertainties at either end are large, as seen in the plot. The total percentage of variance explained by the emulator for cereal is 92%.

Figure 8.3 shows the corresponding plots for cereal when it is irrigated rather than rainfed. Figure 8.3a shows the observed change in yield (LPJmL) and predicted change in yield (emulator). The observed values are mostly within the 95% C.I given by the emulator. Overall, we see that the change in yield of cereal is well predicted for both rainfed and irrigated crops with the majority of points lying within their uncertainty limits. Further tests involved computing cross-validated root mean squared error (RMSE_{CV}) as given by equation (5.73) in Chapter 5. The RMSE_{CV} are 0.0082 and 0.0018 tonne/Ha for rainfed and irrigated respectively. Thus, predictions are much better (low RMSE_{CV}) for



(a) LPJmL and emulator predictions including their 95% C.I vs UN country levels.

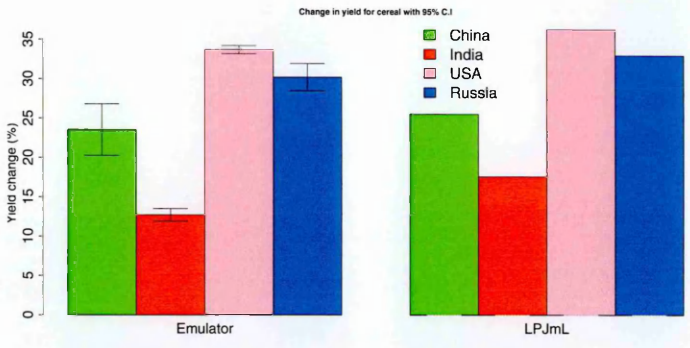


(b) Pair plot of emulator predictions vs LPJmL values including their 95% C.I

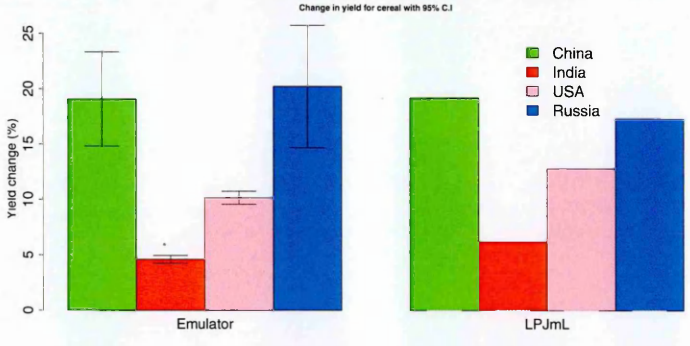
Figure 8.3: Cross-validation for irrigated cereal yield; LPJmL and emulator with its 95% C.I for each countries for mean decadal change in yield between (2085-2094) and (2005-2014) with their 95% C.I. The plots correspond to crop-yield for management level of 5, RCP6 with CO_2 fertilization from UKMO-HADGEM1.

irrigated change in yield than for rainfed change in yield.

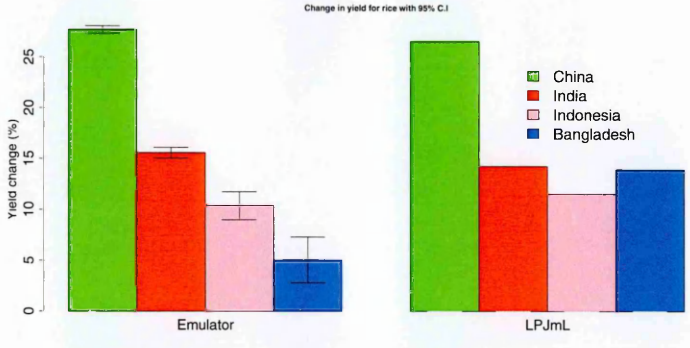
Results for the major producing countries are given in Figures 8.4 and 8.5. The four largest producers of cereal are China, India, USA and Russia and LPJmL



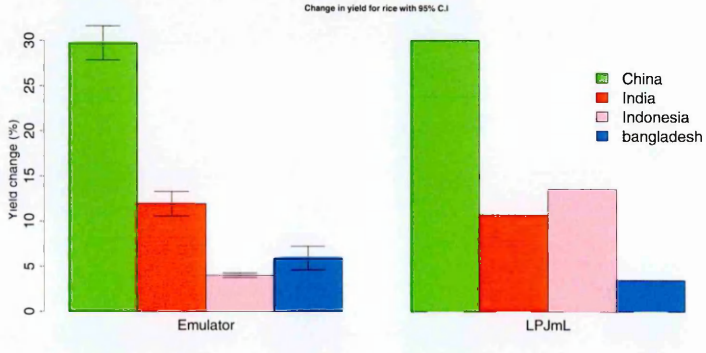
(a) Rainfed cereal



(b) Irrigated cereal



(c) Rainfed rice



(d) Irrigated rice

Figure 8.4: Change in percentage yield by emulator with 95% C.I. in comparison with LPJmL values for major producers of selected crops.

and emulator change in yield of rainfed cereal are shown in Figure 8.4a. Emulator predictions are similar to LPJmL values except for India where the emulator under-predicts the change in yield. (LPJmL gives a 17% change while the emulator predicts a change of $13 \pm 0.4\%$). There is a substantial change in yield of cereal in the USA, larger than for other major producers of cereal crop. For the change in irrigated cereal in Figure 8.4b, the emulator predicts relatively well the change in yield for China, India and Russia. In each of these cases, the LPJmL values are within the 95% C.I band given by the emulator. There is a general increase in yield under climate change in these four countries with a growth of 6-36% across both rainfed and irrigated crops.

Figures 8.4c and Figure 8.4d give the change in yield of rice for its four largest producers. China has a substantial change with a rise of 26% in yield for rainfed and a rise of 30% under irrigation. The emulator predicts $28 \pm 0.18\%$ and $30 \pm 2.2\%$ respectively. The results of other countries are relatively close, except for rainfed rice in Bangladesh where the emulator under-predicts the change in yield. For irrigated rice, the emulator under-predicts the change in Indonesia and over-predicts the change for Bangladesh. In Figure 8.5a, rainfed maize shows a change in yield of slightly more than 2% in China and Brazil but in contrast, there is a relatively steady reduction in yield in USA and France as simulated by LPJmL. The emulator does not capture these changes well. For instance, while LPJmL produces a growth of 2.5% in Brazil, the emulator projects a reduction of $1.2 \pm 0.12\%$. In France, LPJmL gives a decline of 1.5% and the emulator predicts a rise of $4 \pm 1\%$.

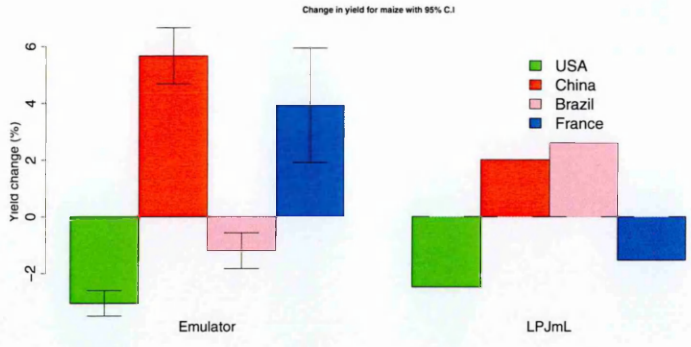
Emulator predictions for irrigated maize in Figure 8.5b are much better than for rainfed maize. LPJmL values are within the 95% confidence band given by the emulator. A slight increment in yield of less than 3% is predicted in China and USA while France has a considerably larger predicted growth of 20%. However, there is a decline in maize yield in Brazil. Figure 8.5c shows a large increase of more than 60% in the yield of rainfed oil crops in Argentina and Figure

8.5d indicates a large increase in irrigated oil crops in three of the four largest producers, Brazil being the exception. Overall, there is an increase in yield for most of the countries. This was expected because of the beneficial effect of CO_2 fertilization, included in this simulation.

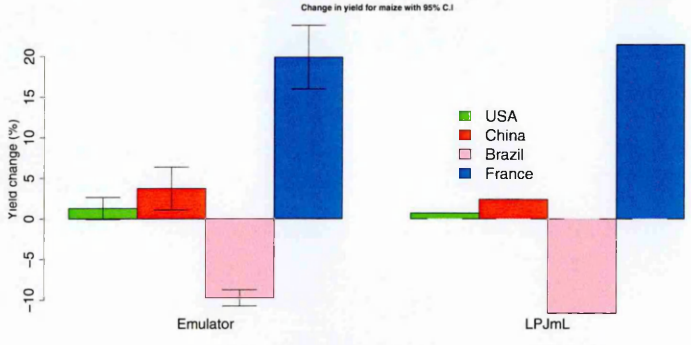
We now consider the spatial map of rainfed cereal in Figure 8.6 (top). The white coloration in this map represents regions that are not currently growing cereal crop. Most of the countries are reasonably well-predicted. Though, the Russia Republic and Canada are under-predicted by the emulator. LPJmL simulates increases in yield of 0.52 and 0.90 tonne/Ha for Russia and Canada, while the emulator predicts growth of 0.49 ± 0.01 and 0.61 ± 0.09 tonne/Ha respectively. Values predicted by the emulator are noticeably higher than the LPJmL values for Mongolia, Tajikistan and South Africa. For instance, while LPJmL gives an increase in yield of 0.20 tonne/Ha for South Africa, the emulator predicts 0.38 ± 0.003 tonne/Ha.

With the beneficial CO_2 effect that we have been considering, both the emulators and LPJmL predict increases in yield for most countries. There are significant yield increases in Europe, China, USA and Russia - countries that all have a well-managed land system. In particular, Wu et al. (2014) associated significant increases in food production in China with technological advances and changes in agronomic practices in that country. In addition, an increase in warming conditions, as projected under climate change, will give rise to a prolonged growing season in these regions which will increase yield (Bates et al., 2008; Stocker, 2013; Zhou et al., 2013).

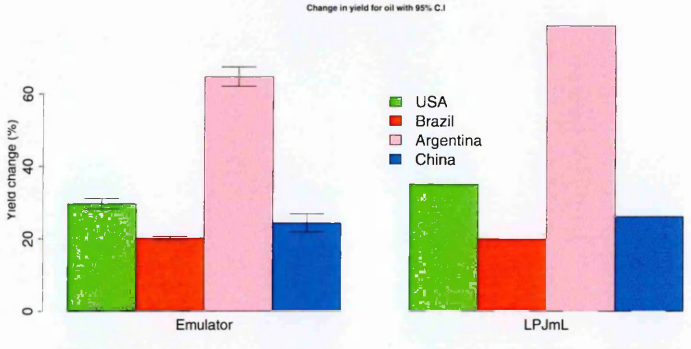
Yield changes of cereal are significantly higher for high latitudes and middle latitudes than for low latitudes, in agreement with the literature (Rosenzweig et al., 1994; IPCC, 2007; Parry et al., 2005; Vermeulen, 2014). Climate change will cause low latitudes to experience a greater degree of heat and water stress, which will often cause a decline in yield even in the presence of the CO_2 fertilization effect. For instance, in Australia there is a low change in yield that could probably



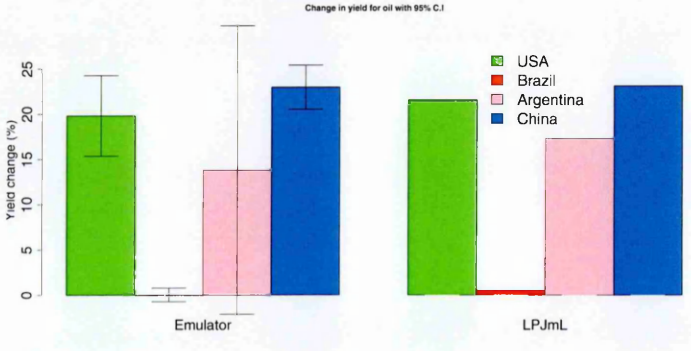
(a) Rainfed maize



(b) Irrigated maize



(c) Rainfed oil



(d) Irrigated oil

Figure 8.5: Change in percentage yield by emulator with 95% C.I. in comparison with LPJmL values for major producers of selected crops.

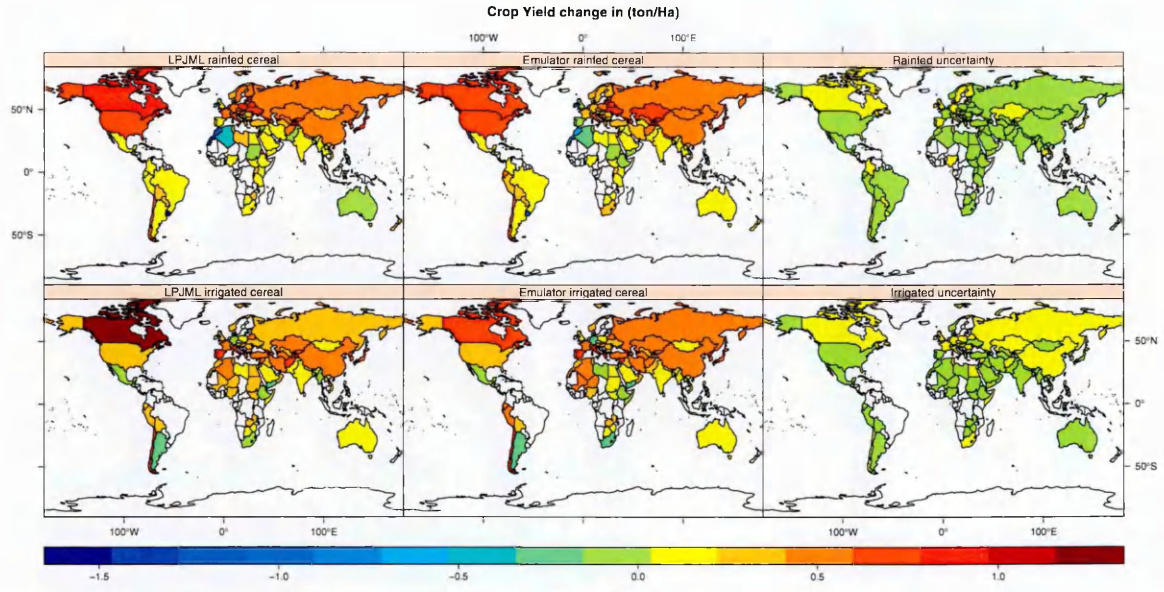


Figure 8.6: Cross-validation of the comparison between LPJmL and LPJmLem for rainfed and irrigated temperate cereals, plotted as mean decadal change in yield between (2085-2094) and (2005-2014) with their 95% C.I. The plots correspond to crop-yield for management level of 5, RCP6 with CO_2 fertilization from UKMO-HADGEM1. The top row plots are for rainfed crops while the bottom row plots are for irrigated crops. The first column is LPJmL values, the second column is emulator predictions, the third column is for emulator confidence interval.

be due to a large variability in precipitation, extended droughts and a greater incidence of extreme weather events that could disrupt agricultural production. In addition, Australia and New Zealand exhibit a wide diversity of climates with a significant constraint on water resources as described in Stocker (2013). Algeria, Ireland, Morocco and Uruguay also have a substantial predicted reduction in yield. Uncertainty levels are low and vary from one country to another because each country is modelled by an independent Gaussian process regression.

Under irrigation, Figure 8.6 (bottom) shows the change in yield for cereal in USA, China and Russia is lower than when the crop is rainfed, unlike Canada that has a larger increase in yield when the crop is irrigated. The reduction in crop yield is larger in Argentina, Yemen, South Africa and Ethiopia. The increase in cereal yield is substantial for Canada, Armenia and Spain.

Next, we extend our cross-validation to other crops. Figure 8.7 (top) is for rainfed rice. The plot indicates that the emulator predicts the change in yield

reasonably well in most countries, though the emulator under-predicts the change in yield in Russia and Argentina. Both LPJmL and the emulator produce a higher increase in yield for the Russia Federation and Argentina than for the USA and China. The uncertainty around predictions in most countries is quite low.

For irrigated rice (bottom) plot, LPJmL values are quite well predicted in most countries. Chile, Morocco, Macedonia and Hungary have a substantial increases in yield, more than in other countries. Uncertainty levels are quite low for most of the predictions. In Figure 8.8 (top) for example, predictions are quite close to the LPJmL values in most regions for both rainfed and irrigated crops. The uncertainty associated with predictions is fairly similar in most countries, and the values are small. The Russia Federation has a high value under rainfed rice as we have earlier observed, and a similar high value is also apparent for the irrigated crop. Maize, unlike other crops, exhibits little change in yield under this scenario.

Figure 8.9 for rainfed oil shows a rise in yield for most countries, with emulator predictions comparable to the LPJmL simulation. The results are also similar under irrigation and uncertainty levels are generally low.

We summarise the performance of the emulators in Table 8.2, which calculates the overall proportion of the variation in LPJmL prediction for the cross-validated data that is explained by the emulator. The results are over all management levels, RCP and time point with CO_2 fertilization for UKMO-HADGEM1. We also compare the performance of the GP emulator with the WLS techniques used in Chapter 7. In this case, we use the same procedure set out in this chapter, but instead of performing GP regression we used the WLS approach. The proportions of variance explained when WLS is used is also shown below in Table 8.2.

Results for the GP regression are slightly better than with the WLS method for all the cases except for rainfed oil and irrigated rice. Results for irrigated cereal and maize are much better under the GP models than with the WLS approach. The two techniques produce relatively similar results for irrigated rice and oil.

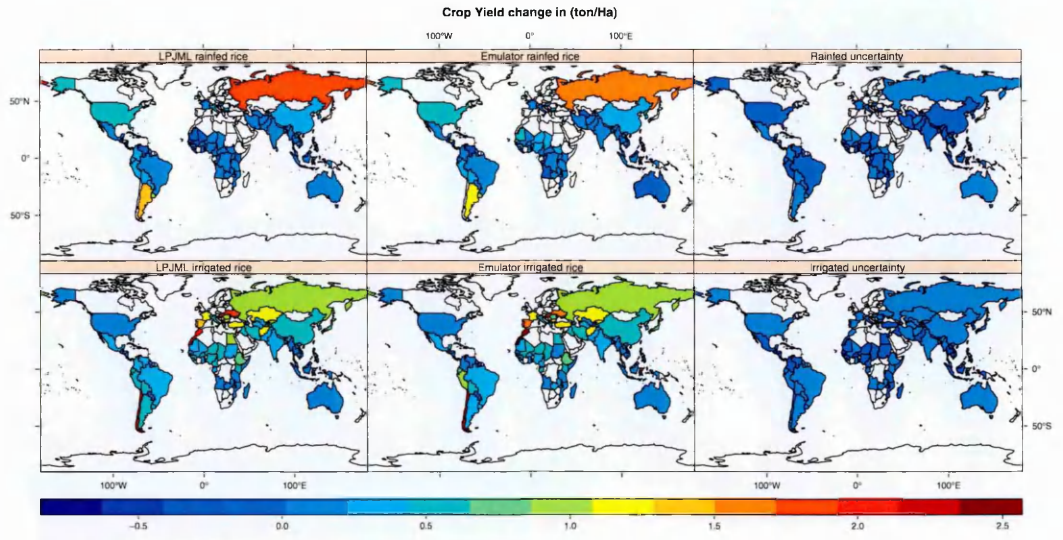


Figure 8.7: Cross-validation of the comparison between LPJmL and LPJmLem for rainfed and irrigated temperate rice, plotted as mean decadal change in yield between (2085-2094) and (2005-2014) with their 95% C.I. The plots correspond to crop-yield for management level of 5, RCP6 with CO_2 fertilization from UKMO-HADGEM1. The top row plots are for rainfed crops while the bottom row plots are for irrigated crops. The first column is LPJmL values, the second column is emulator predictions, the third column is for emulator confidence interval.

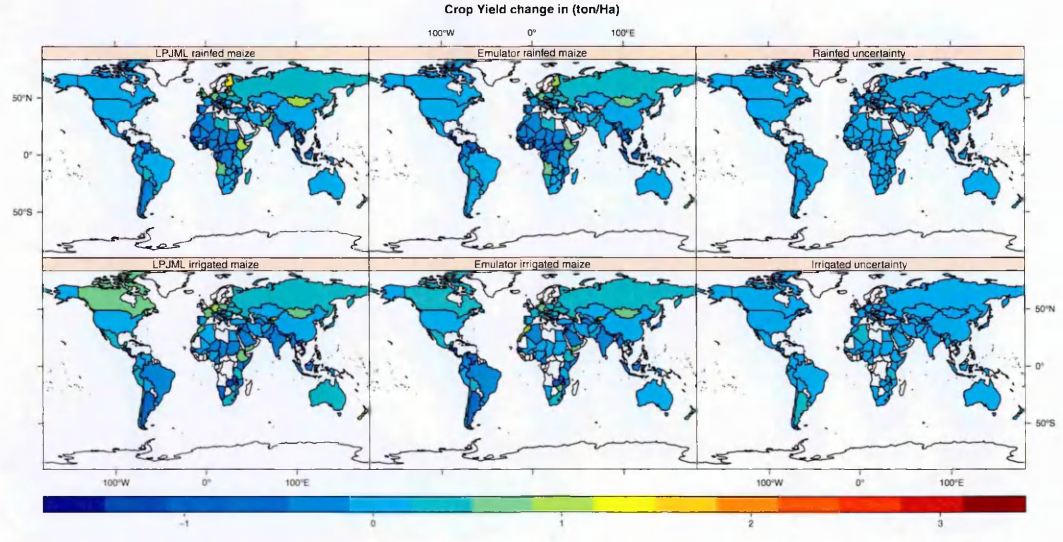


Figure 8.8: Cross-validation of the comparison between LPJmL and LPJmLem for rainfed and irrigated temperate maize, plotted as mean decadal change in yield between (2085-2094) and (2005-2014) with their 95% C.I. The plots correspond to crop-yield for management level of 5, RCP6 with CO_2 fertilization from UKMO-HADGEM1. The top row plots are for rainfed crops while the bottom row plots are for irrigated crops. The first column is LPJmL maps, the second column is emulator maps, the third column is for emulator confidence interval.

Overall, the GP emulator explains between 70-94% variance while WLS ranges from 63-93% for both rainfed and irrigated crops.

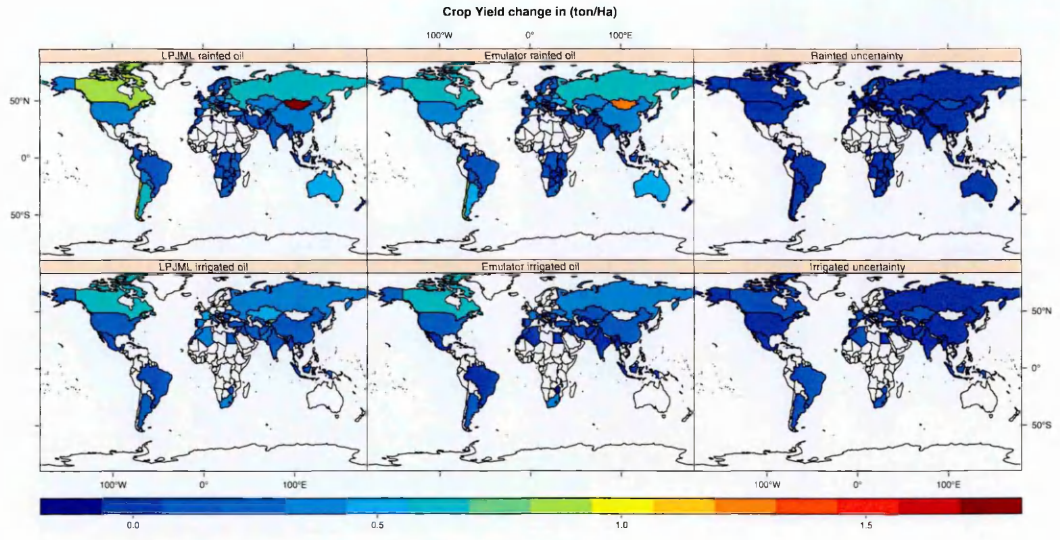


Figure 8.9: Cross-validation of the comparison between LPJmL and LPJmLem for rainfed and irrigated temperate oil, plotted as mean decadal change in yield between (2085-2094) and (2005-2014) with their 95% C.I. The plots correspond to crop-yield for management level of 5, RCP6 with CO_2 fertilization from UKMO-HADGEM1. The top row plots are for rainfed crops while the bottom row plots are for irrigated crops. The first column is LPJmL maps, the second column is emulator maps, the third column is for emulator confidence interval.

Results from the cross-validation with four additional GCMs are given in Table 8.3. We see that results for CCSR-MIROCMED are much better than for other GCMs. This is because CCSR-MIROCMED is very similar to CCSR-MIROC32HI, which gave part of the training data. Results from GP models are better than from the WLS technique among all GCMs except for rice under IPSL-CM4, where the result is slightly better than for GP. The improvement give by the GP model over WLS is often substantial.

Table 8.2: Table of cross-validated proportion of variance ρ showing the overall performance of the GP emulators, and a comparison with the WLS method, for rainfed and irrigated crops, with all management levels, RCPs, and time slices, but with CO_2 fertilization only. The list of countries are shown in Table 8.1.

ρ	Number of countries		GP		WLS	
crops	rainfed	irrigated	rainfed	irrigated	rainfed	irrigated
cereal	106	83	0.89	0.94	0.81	0.87
rice	79	95	0.82	0.92	0.77	0.93
maize	141	100	0.70	0.80	0.63	0.76
oil_{max}	139	64	0.89	0.88	0.91	0.88

Figures 8.10a shows the average global changes in yield for rainfed crops that

Table 8.3: Comparison of GP with WLS methods for four GCMs, with CO_2 fertilization, management level 5, RCPs 4.5 and 8.5, and all time slices for rainfed crops. The values are the proportion of variance ρ explained.

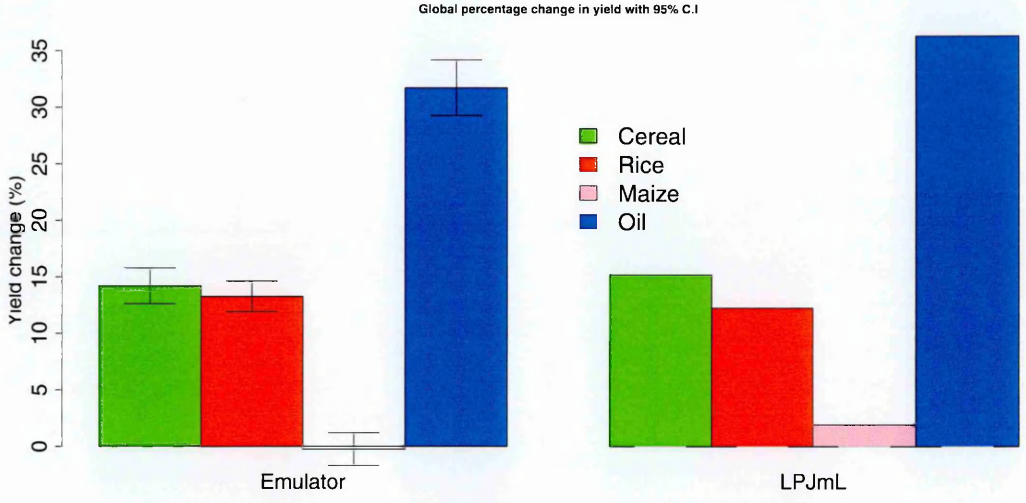
Crop	CCSR-MIROCMED		GISS-MODELER		GISS-MODELEH		IPSL-CM4	
	GP	WLS	GP	WLS	GP	WLS	GP	WLS
Cereal	0.87	0.81	0.71	0.66	0.78	0.72	0.75	0.72
Rice	0.83	0.81	0.71	0.67	0.83	0.67	0.71	0.72
Maize	0.92	0.86	0.69	0.67	0.76	0.68	0.87	0.80
Oil[1]	0.89	0.83	0.84	0.78	0.83	0.74	0.83	0.77

[1] Oil=yield$_{max}$[soybean, rapeseed, sunflower].

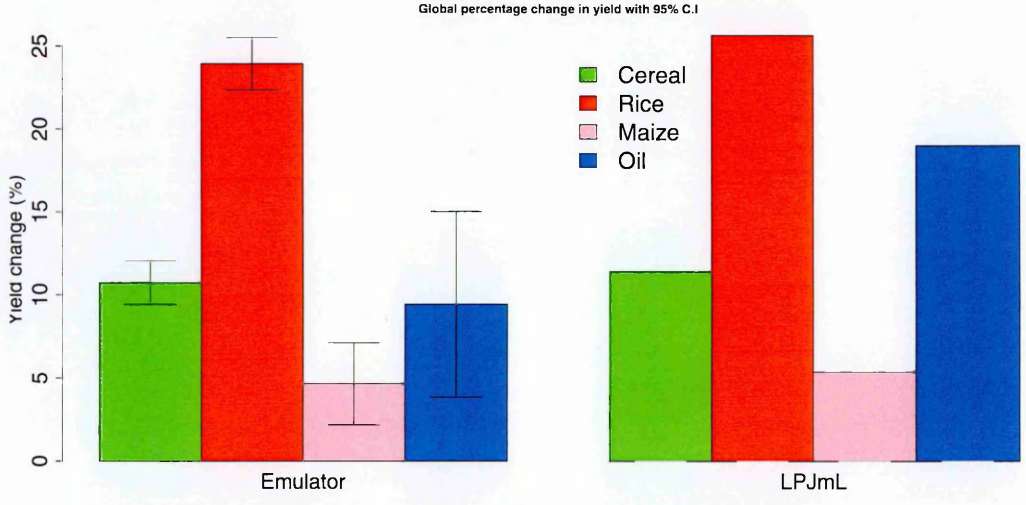
are predicted by the GP emulator in all countries and compares them with LPJmL values. The emulator predicts the global average change in rainfed yield for cereal and rice relatively well and LPJmL values are within the 95% confidence interval. However, rainfed yields of the maize and oil crops are under-predicted by the emulator. LPJmL simulates a small increase of 1.9% for maize; the emulator predicts a reduction of -0.25%. Figure 8.10b for irrigated crops indicates that the yield from the oil crops is under-predicted by the emulator, while emulator predictions for cereal, rice and maize are relatively similar to the LPJmL values. Uncertainty is much higher for the yield from oil crops than for other crops. Comparing the top and bottom plots in Figure 8.10b, for irrigated rice and maize, the global percentage changes in yield are higher than for rainfed rice and rainfed maize. In contrast, for irrigated cereal and oil the percentage changes are a little lower than for the rainfed crops.

8.5 Conclusion

In this chapter, we have demonstrated a means of making inference about the parameters of the emulator using GP regressions. Combinations of OLS, PCA and GP methods were used to emulate major crop-yield as a linear function of seasonal



(a) Rainfed crops



(b) Irrigated crops

Figure 8.10: Global percentage change in yield by emulator with its 95% C.I. in comparison with LPJmL values for a period (2085-2094) relative to (2005-2014) with their 95% C.I. The plots correspond to crop-yield for management level of 5, RCP6 with CO_2 fertilization from UKMO-HADGEM1.

climate variables and other relevant variables. Aggregated country average of actual yield change was calculated by combining yields simulated by LPJmL for both rainfed and irrigated crops. We consider eight time-slices averaged over ten years with the baseline corresponding to the average in the decade (2005-2014).

In order to implement GP, it was necessary to reduce the computational burden by aggregating the OLS residuals from a fine to a more coarse resolution on a country level. GP is computationally intensive, making it important to reduce

the dimensionality of data for tractable application of GP modelling. From a methodological viewpoint, the most important finding is that the GP regression appears to be better than WLS as a means of performing stage 2 analysis. The cross-validation results (Table 8.2) found that the GP emulator explained 70-94% of the variance in LPJmL values, while WLS explained only 63-93%. Hence, while GP cannot be applied at a grid-cell level (too many regression would be needed), it would seem to be the preferable for stage 2 when it is viable to use it.

Another advantage of the GP method is that the prediction is probabilistic so confidence intervals can be formed as a measure of uncertainty. Also GP has some flexibility because different correlation and linear regression functions can be specified. Additional information in the form of a prior distribution could, in principle, be used to improve the prediction.

These emulators can predict the change in global crop yield as a function of climate from any GCM/RCP combination and for different management level options and CO_2 fertilization level. Overall, cereal is more predictable in South America, Africa and China than in other parts of the world. Similarly, rainfed rice is more predictable in North America, China and India but is more difficult to predict in Australia. Maize and oil can be predicted well in most countries, Canada and Mongolia values being exceptions. Uncertainty levels are also relatively low for all the four crops and there are only a few countries with wide confidence band, which could be attributed to a large variability in the training data for those countries.

Chapter 9

Summary

The introduction chapter gave the scope of work in this thesis as the use of statistical emulation as a cheaper surrogate to study the impact of climate change on terrestrial ecosystems. In this chapter, we will conclude by describing the progress made so far towards achieving that stated objective. Suggestion for some possible future research that could emanate from this study and towards a wider application on integrated impact assessments will be provided.

9.1 Discussion

We presented a statistical method for emulating the underlying physical dynamics of carbon fluxes and crop-yield responses to change in climate. This thesis has addressed the joint emulation of the impact of climate change; CO_2 fertilization effect and crop management levels on global crop yields. In addition, we have emulated NPP, FC and HR using a variety of techniques. We applied a combination of OLS, WLS, PCA, CR and GP regressions to fit our models to the response data. The emulators are designed to predict the change in global crop yield and carbon fluxes as a function of climate and other important variables like soil, LAI, CO_2.

In Chapter 1, we introduced the work to be covered in this thesis and its motivation. Chapter 2 dealt with general background. We reviewed literature

that relevant to the research work to be carried out in Chapter 3, while Chapter 4 provided descriptions of important models that simulated the data we used in this study. In Chapter 5, we described the different statistical methodologies used for the analyses. Chapters 6, 7 and 8 developed and tested novel approaches to statistical emulation and we then obtain important results. The results are novel contributions to the knowledge about the impact of future climate change on global vegetation. We provided a statistical emulation procedure as a viable and cheaper alternative to the process-based simulation model of LPJmL that is computationally expensive to run in the assessment of global impact of climate change on vegetation.

9.2 Conclusion

This thesis combined several statistical techniques in new ways to form effective emulations of global carbon fluxes and crop yields. Chapter 6 described the emulation of change in carbon fluxes for NPP, HR and FC. Chapters 7 and 8 developed a procedure for constructing an emulator for LPJmL simulations of potential crop yield for cereal, rice, maize, groundnut and oil crop functional types. Two emulators were built for each crop, one for the rainfed crop and the other for its yield under irrigation. Each emulator was constructed using a novel form of two-stage process. The first stage used OLS to fit crop-yield as a smooth function of climate variables under the assumption that each spatial point is an independent sample. The second stage involves interpolation of the spatial residuals of unknown scenarios from known scenarios (an approach similar to pattern scaling) using a combination of WLS and PCA. We made similar assumptions to those of GP emulators (O'Hagan, 2006; Oakley & O'Hagan, 2004; Conti et al., 2008). In particular, it is assumed that the emulator is a smooth and continuous function of its input variables. The second stage indeed improves the overall predictions.

We used cross-validation to test the accuracy of the emulator and performed a

sensitivity analysis for each crop response. LPJmL uses daily time-steps as crop-yields respond to daily variability. Here, we have chosen to use only seasonal mean climatic variables as input as these are readily available input data. With these inputs, under cross-validation the emulators explained 62-93% of the variance for irrigated crops and 60-88% of the variance for rainfed crops. Sensitivity analysis indicated that the predicted yield of rainfed crops depends most heavily on baseline seasonal temperature, CO_2 change, latitude and LAI. Irrigated crops are dominantly sensitive to temperature and less dependent on precipitation, as expected. We provided spatial plots visually to compare the performance of the emulators with the LPJmL simulations.

We noted that using Bayesian technique of GP emulator directly on LPJmL data would be challenging, because of the large number of parameters to be estimated. We have to sample from all scenarios (time-slice, RCP, GCM, management levels). In order to overcome the problem, it was necessary for us to reduce the resolution and aggregate data to a country level in order to reduce the computational load. GP was used to emulate these data because of its wider applicability and flexibility in modelling complex data. Although Bayesian methods require much experience in selecting prior distribution of hyperparameters, this was not a problem in this study as we used conjugate priors. We applied a combination of PCA-GP in Chapter 8 to perform GP regression of crop yields. Cross-validation under GP regression in Chapter 8 explained between 55-86% and 60-89% of the variance of the response data.

LPJmL crop data are characterized by a large proportion of zero observations. Grid points where particular crops are not currently grown are represented as zeros in the simulation. A censored regression approach (Moore et al., 2000; Cai & Cheng, 2004) was also used to model these data by treating the zero observations as censored observations (results shown in Chapter 7). However, this method was not helpful because of the large dataset we have, coupled with the fact that the censoring algorithm takes a longer computation time, even after

reducing the data size from $0.5° \times 0.5°$ to $2° \times 2°$ resolution.

Nevertheless, for the first stage we analysed a small sample of this data with censored regression but the results were not better than OLS. Using either non-linear regression or a Gaussian process emulator would be challenging because of the large number of parameters to be estimated with samples from many scenarios (time-slice, RCP, GCM, management levels, irrigation regime).

We might also have used a dynamic emulation by emulating each grid-cell individually as a function of time. Rather, our choice of OLS is driven by its simplicity.

Our crop emulation approach extends the work of Lobell & Burke (2010, 2008), which models temporal and spatial variation to predict future crop yields from climate variables and performs sensitivity analyses to examine the importance of temperature and precipitation on future yields. However, unlike Lobell & Burke (2010, 2008), our predictions were not based solely on climate variables but also incorporated soil, latitude, crop management levels and other covariates. Our analyses and aims are also broader; Lobell & Burke (2010) work with just 94 crop-region combinations and only examine temperature and precipitation while Lobell & Burke (2008) work solely with the yield of maize in 200 sites in Sub-Saharan Africa. Our emulators provide estimates of projected change in crop yields at any level of CO_2 emission on high spatial gridded resolutions for five different crops. We also performed rigorous and extensive cross-validation for several climate models and RCPs, and our global sensitivity analysis measures the individual contribution of a number of different variables to the overall uncertainty. In agreement with Lobell & Burke (2008), the clearest result from our sensitivity results is that temperature is the dominant source of uncertainty in future impacts assessment. The sensitivity analysis used a variance based decomposition technique and the contributions of other variables that it quantifies have not previously been reported.

The emulators reduced the LPJmL model down to a two-stage process that is

capable of predicting global crop-yields of different crop functional types, and the spatial distribution and temporal dynamics of these yields in response to a changing climate. These emulators are much faster to run compared to the LPJmL model. LPJmL is computationally expensive to run, while these emulators give results almost instantaneously. The LPJmL emulator, without considering the MAGICC and ClimGen simulation time, takes about $8 - 10$ minutes to produce eight decadal changes of crop-yield data (on $0.5° \times 0.5°$ resolutions) with a 24G RAM, 4-cores Window machine. This approximately 60-fold increase in computational efficiency is particularly useful when the model is coupled to one or more models for the integrated assessment of climate impact.

There will be limitation in using an emulator as a substitute for LPJmL because emulator outputs will only be an approximation to the LPJmL. Thus, there will be some differences (error) between outcomes predicted by the emulator and outcomes observed from LPJmL. The major objectives of this thesis are to obtain good predictions of the major simulation outputs from LPJmL and to provide a measure of uncertainty associated with these predictions.

This thesis described statistical emulators that are useful to approximate complex models when the simulator is time-consuming due to the computational expense of the model. These emulators provide good approximations to the true dynamics of LPJmL response outputs to climate and CO_2 emissions. In addition, these emulators have been applied already as a flexible way for data exchange among the coupled models that were used in the ERMITAGE project. In particular, it has been coupled with the GEMINI-E3 model in the ERMITAGE project.

The ERMITAGE project in general provided useful information for decision and policy-makers on the interactions between climate, agricultural, ecosystems, economy and technologies in order to effectively assess the economic impact of global climate change.

9.3 Future work

This thesis has demonstrated emulation techniques as a means for assessing the impact of climate change on global terrestrial vegetation. Several techniques of emulating global carbon fluxes and crop yields have been used. However, there are still some interesting analyses we could not cover due to time constraints and non-availability of required simulation data.

Throughout this thesis, we have applied a decadal averaging as a strategy to condense the dimension of the data. It is not clear as at now the effect of this decadal averaging on the emulator. It will be interesting to assess the implication of decadal averaging on the emulator algorithm using another data reduction strategy that can minimize non-relevant information. Similarly, we have used various regression techniques namely OLS, WLS, censored and GP regression for our parameter estimation. Alternative approaches for estimating the model parameters in this thesis could also be further investigated.

The use of LPJmL emulator to interpolate paleoclimate data is another immediate and beneficial application of the work in this thesis. Examination of past climates and the forcing that caused them to change will provide a better understanding of how future climate will change. There is currently little simulation data on crop yields for paleoclimate study and it is difficult to run a simulation model for the long times scales associated with paleoclimate studies. Emulation is a useful tool for making such projection.

The line of research that follows from Chapter 8 is a Gaussian process regression that is flexible and widely used in various applications. In this thesis, we have used an exponential correlation function for the GP to model an unknown function emanating from the residuals of change in crop yield. Examination of the sensitivity of the GP model to the choice of prior distribution parameters and investigation of robustness to different choices of correlation functions could potentially improve the results in this thesis.

We have investigated the contribution of each climate variable to the overall

variance in the LPJmL output through sensitivity analysis. However, LPJmL has a large number of parameters that are calibrated for the simulation of outputs. A quantification of LPJmL parametric uncertainty by considering the sensitivity of the output to the individual parameters will enhance the understanding of the most influential parameters and improve the interpretation of our sensitivity analysis. An improved understanding of relevant parameters is essential for making a better choice in designing plausible experimental points. In addition, it will help to determine the influence of parametric uncertainty on the LPJmL simulation outputs. These analyses are not performed in this work because there are no available simulation data.

There is a need to assess the full range of local climate changes, considering particularly the uncertainty stemming from the often substantial differences in precipitation projections among GCMs (e.g. Knutti & Sedlacek (2013); Ramesh & Goswami (2014)). This thesis focused only on few selected GCMs to demonstrate the impact of climate change on vegetation rather than attempting to provide a comprehensive assessment based on projections from the $\sim$ 20 GCMs available in global climate data archives such as CMIP3 or CMIP5. It is possible for the crop and carbon flux response patterns to differ under different climatic projections from other GCMs training data. Further research to incorporate a large number of GCMs will enhance the findings in this study.

For a broad understanding of the impacts of climate change on agriculture, quantification of potential impacts on most of the agricultural crops are essential. In order to address this issue, we have adopted an emulation approach for studying the impact of climate change on some selected agricultural crops (cereal, rice, maize, groundnut and oil (sunflower, rapeseed, soybean)) but there are still a large number of other crops (including biofuels) that could also offer useful information on the impact of climate change on global food production and energy security. Another possible extension of this study is to accommodate more crop functional types in the modelling to provide a comprehensive assessment.

Another interesting aspect that we could have covered in the thesis is the assessment of climate change impact on water runoff. Evaluation of the potential impact of climate change on water resources and its availability is essential. Water is valuable for its supporting role in ecosystems function and irrigated agriculture. Higher temperatures and increased variability of precipitation could lead to an increased irrigation water demand. The world population is projected to increase by the middle of this century (Crossette, 2010). This increase will cause a rise in demand for major food crops thereby causing a growing pressure on water resources. The problem will worsen under climate change because of rising temperature and changing precipitation patterns that will affect crop production (Parry et al., 2004). Change in precipitation patterns affects water availability and runoff. Relatively few studies have investigated the impacts of climate change on the hydrology cycle and runoff and studies that examine projection of natural water cycle variability are limited by inter GCM uncertainty.

There is a rise in demand for projections about the potential impacts of climate change on water resources from decision-makers. A better understanding of projected climatic impact on the hydrological cycle is essential. A sound knowledge of the hydrological cycle will enhance better management of global water resource. Water-climate studies are necessary to understand the major sources of uncertainty in water modelling. Further work is needed to explore the impact of climate change on a simulated water runoff from LPJmL, which can also be addressed using a statistical emulation. Emulation offers a cheaper and relatively quicker alternative approach and can quickly evaluate a large number of scenarios for a policy-maker.

Appendix A

List of emulators

(i) NPP emulator

(ii) Fire carbon emulator

(iii) Heterotrophic respiration emulator

(iv) Temperate cereal emulator (rainfed)

(v) Rice emulator (rainfed)

(vi) Maize emulator (rainfed)

(vii) Oil-crop emulator (rainfed)

(viii) Cereal emulator (rainfed, high resolution)

(ix) Rice emulator (rainfed, high resolution)

(x) Maize emulator (rainfed, high resolution)

(xi) Oil-crop emulator (rainfed and irrigated, high resolution)

(xii) Cereal emulator (irrigated, high resolution)

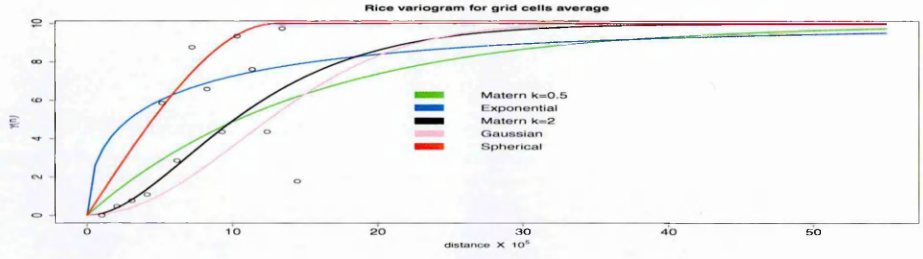
(xiii) Rice emulator (irrigated high resolution)

(xiv) Maize emulator (irrigated, high resolution)

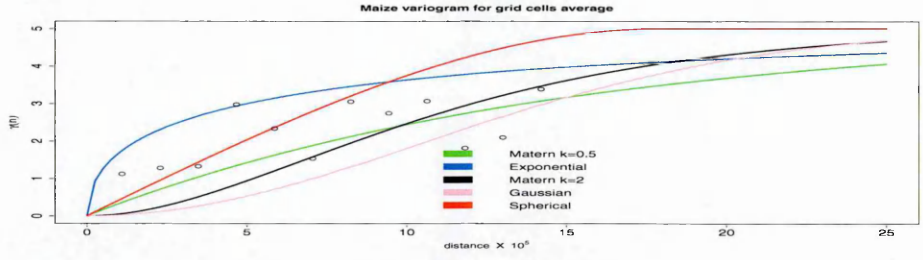
(xv) Groundnut emulator (irrigated and rainfed, high resolution)

Appendix B

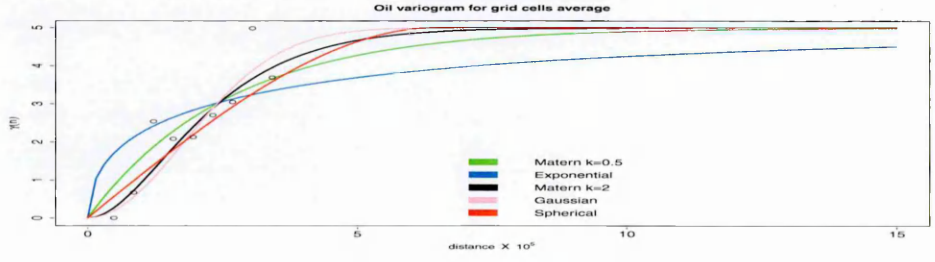
Variogram plots



(a) Rainfed rice

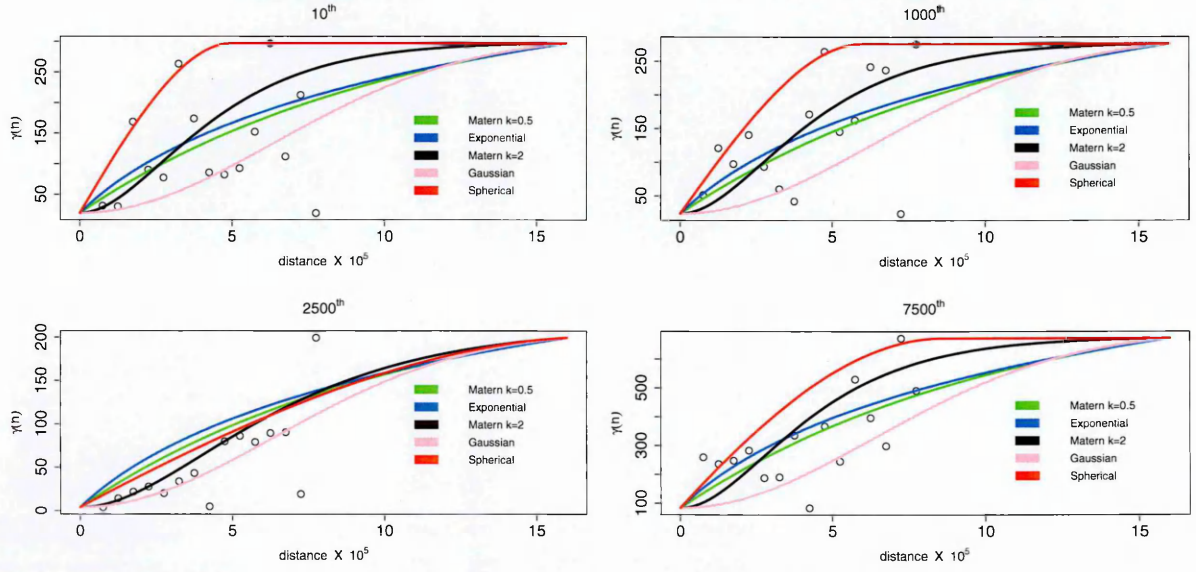


(b) Rainfed maize

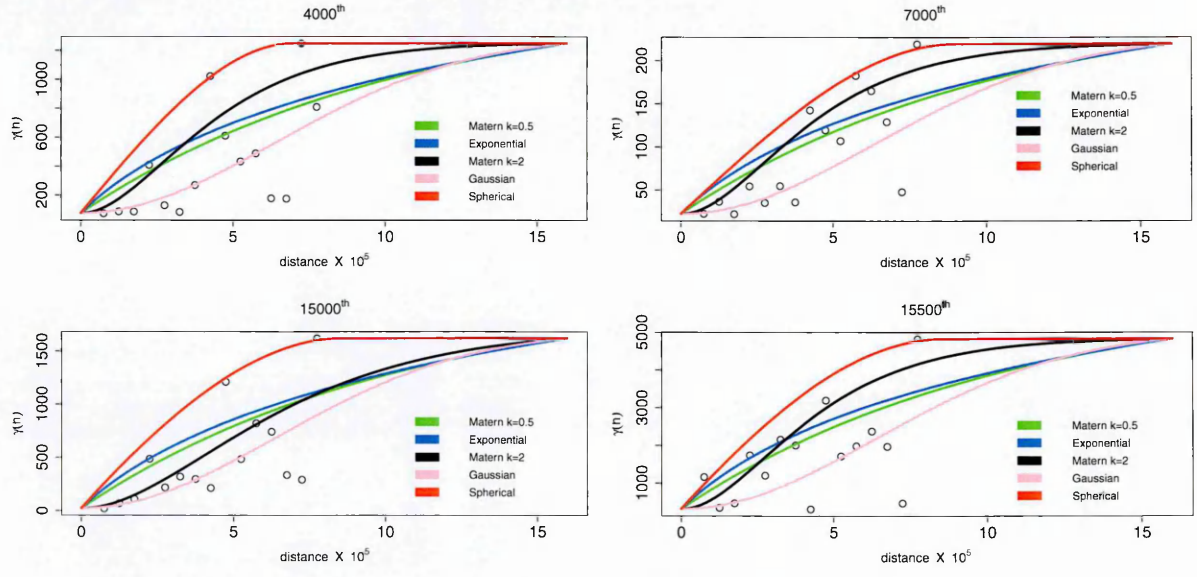


(c) Rainfed oil

Figure B.1: A sample of further variograms that again show an increase in $\hat{\gamma}(\overline{d}_\ell)$ as $\overline{d}_\ell$ increases for rainfed rice, maize and oil averaged over all grid cells. The points are the estimated variogram bins from the residual data while the curves are the theoretical models fitted using the covariance functions.



(a) rainfed cereal at $N_k^{th} = 10^{th}, 1000^{th}, 2500^{th}, 7500^{th}$ grid points

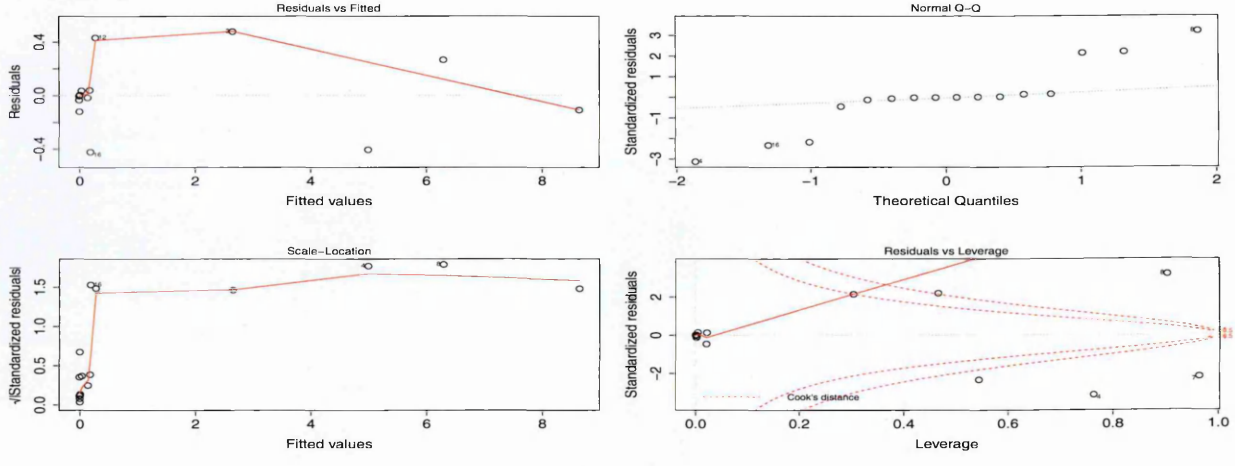


(b) rainfed cereal at $N_k^{th} = 4000^{th}, 7000^{th}, 15000^{th}, 15500^{th}$ grid points

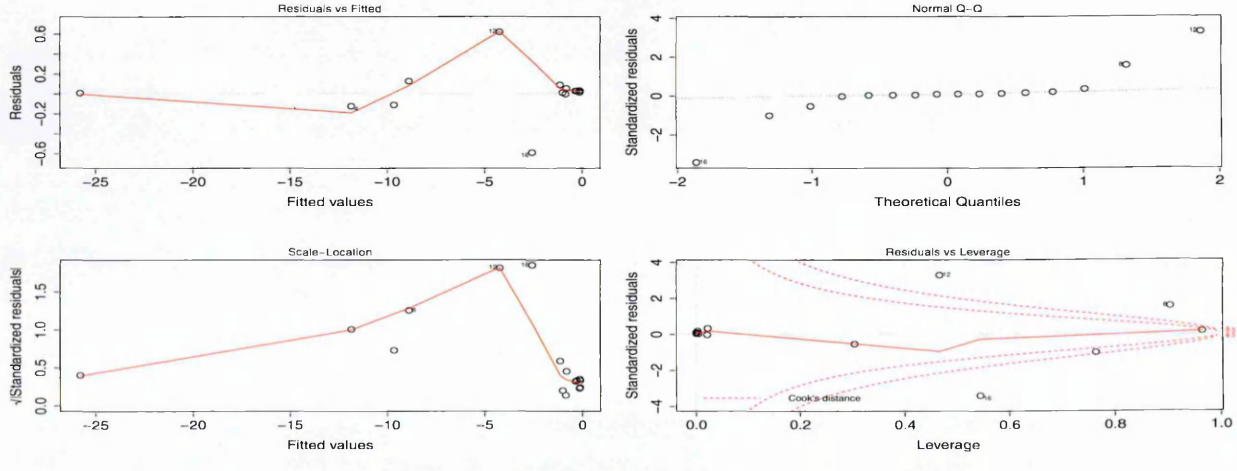
Figure B.2: A sample of further variograms that again show an increase in $\hat{\gamma}(\overline{d}_\ell)$ as $\overline{d}_\ell$ increases for some randomly selected grid points under rainfed cereal.

Appendix C

WLS diagnostics

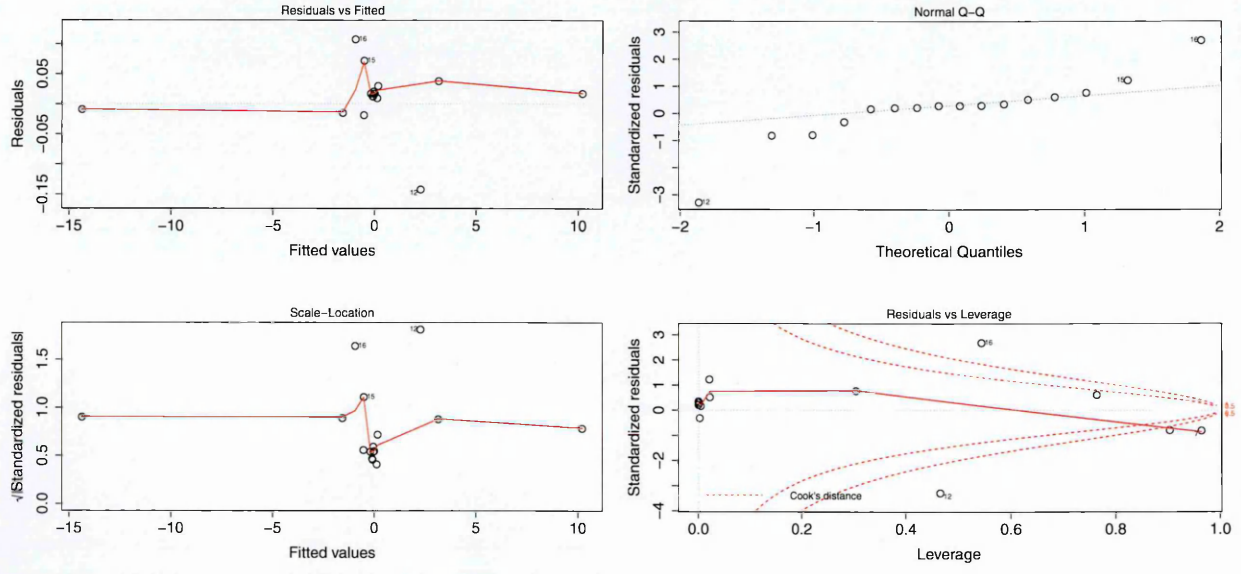


(a) cereal at 7000^{th} grid point

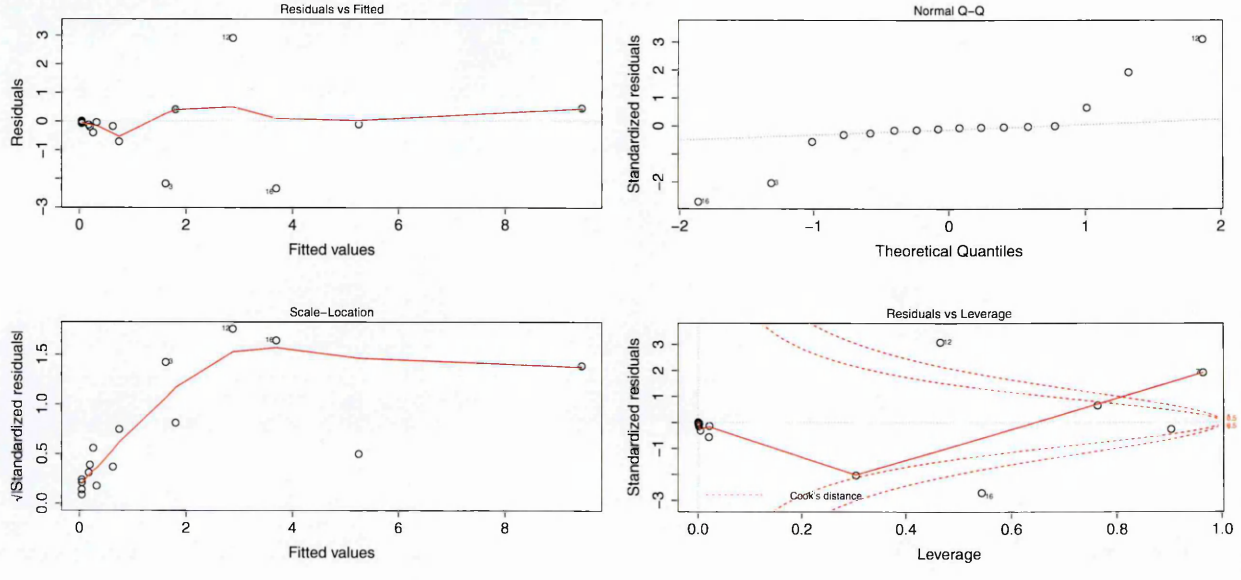


(b) cereal at 14000^{th} grid point

Figure C.1: WLS diagnostic plots for some randomly chosen grid points for rainfed cereal. The points are the 16 data values representing scenarios.



(a) cereal at 9500^{th} grid point



(b) cereal at 4500^{th} grid point

Figure C.2: WLS diagnostic plots for some randomly chosen grid points for rainfed cereal. The points are the 16 data values representing scenarios.

Appendix D

List of input variables

Table D.1: The emulator's input variables.

Variables	Full names
	1991-2000 data only
csdtr/cwdtr/cspdtr/cadtr	Change in mean diurnal temp range in summer/winter/spring/autumn
csvap/cwvap/cspvap/cavap	Change in mean vapour pressure in summer/winter/spring/autumn
cstmn/cwtmn/csptmn/catmn	mean near surface temp min in summer/winter/spring/autumn
cstmx/cwtmx/csptmx/catmx	mean near surface temp max in summer/winter/spring/autumn
sdtr/wdtr/spdtr/adtr	baseline mean diurnal temp range in summer/winter/spring/autumn
svap/wvap/spvap/avap	baseline mean vapour pressure in summer/winter/spring/autumn
stmn/wtmn/sptmn/atmn	baseline mean near surface temp min in summer/winter/spring/autumn
stmx/wtmx/sptmx/atmx	baseline mean near surface temp max in summer/winter/spring/autumn
	2001-2100 data
scld/weld/spcld/acld	Change in mean cloud cover in summer/winter/spring/autumn
spre/wpre/sppre/apre	Change in mean precipitation in summer/winter/spring/autumn
stmp/wtmp/sptmp/atmp	Change in mean temperature in summer/winter/spring/autumn
swet/wwet/spwet/awet	Mean change in wet day frequency in summer/winter/spring/autumn
iscld/iweld/ispcld/iacld	Initial (baseline) mean cloud cover in summer/winter/spring/autumn
ispre/iwpre/isppre/iapre	Initial mean precipitation in summer/winter/spring/autumn
istmp/iwtmp/isptmp/iatmp	Initial mean temperature in summer/winter/spring/autumn
iswet/iwwet/ispwet/iawet	Initial wet day frequency in summer/winter/spring/autumn
co2 and cco2	Baseline mean CO_2 and change in mean CO_2
soil, lat and LAI	Soil and latitude parameters, and maximum leaf area index

In the northern hemisphere, *summer* = {June July August}, *winter* = {December January February}, *spring* = {March April May} and *autumn* = {September October November}; obvious changes are made for the southern hemisphere.

Appendix E

Glossary

Table E.1: Glossary

Glossary	Full meaning
Emulator	Approximation to simulator which gives predictions of simulator outputs at untried input points.
GCMs, AOGCMs	General Circulation (or Global Climate) Model, Atmosphere-Ocean GCM.
Grid	Horizontal and vertical spatial resolutions.
Projection	A climate prediction along a specified scenario.
RCPs	Representative Concentration Pathways.
Spin-up period	Time taken between the initial time of a climate simulation and the moment when the model reaches its own equilibrium.
Weather	State of the atmosphere (temperature, precipitation, cloudiness)
Climate	Accumulation of daily and seasonal weather events over a long period of time.
Cloud cover	Amount of the sky obscured by clouds when observed at a particular location.
Insolation	Solar radiation that is reflected by atmospheric components to the Earth's surface.
Evapotranspiration	Vaporization of water through direct evaporation from wet surfaces.
Growing degree days	A growing degree-day is defined as a day on which the mean daily temperature is one degree above the base temperature-minimum temperature required for growth of a particular crop.
Growing season	The number of days between last spring freeze date and first fall freeze date.
Precipitation	Any form of water particles-liquid or solid-that falls from the atmosphere.
MAGICC	Model for the Assessment of Greenhouse Gas Induced Climate Change (MAGICC).
MAGICC6	Updated version of MAGICC
ClimGen	spatial climate scenario generator
IPCC	The Intergovernmental Panel on Climate Change.
PFTs	Plant Functional Types (natural vegetation at the level of biomes).
Rainfed crops	Water requirements and water consumption depend on precipitation.
Irrigated crops	Water requirements and water consumption depend on irrigation.
LAI	Leaf Area Index
CCCMA-CGCM31	Canadian Centre for Climate Modelling and Analysis Global Climate Model.
CCSR-MIROC2HI	Center for Climate System Research Model for Interdisciplinary Research on Climate.
UKMO-HADGEM1	Hadley Centre Global Environmental Model, Met Office UK.

References

Abraham, G., & Inouye, M. (2014). Fast Principal Component Analysis of Large-Scale Genome-Wide Data. *PloS one*, 9(4), *e*93766. 68

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F, editors. Second International Symposium on Information Theory. Budapest (Hungary): *Akademiai Kiado*, 267 – 281. 56

Amemiya, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica*, 41(6), 997 – 1016. 38

Andreae, M.O. (1991). Biomass burning: its history, use, and distribution and its impact on environmental quality and global climate change. *Global biomass burning: atmospheric, climatic, and biospheric implications (ed. by J.S. Levine)*, 3 – 21. MIT Press, Cambridge, Massachusetts. 14

Annan, J.D. et al. (2005). Parameter estimation in an intermediate complexity Earth system model using an ensemble Kalman filter. *Ocean Modelling*, 8, 135 – 154. 35

Arnell, N. W., Lowe, J. A., Brown, S., Gosling, S. N., Gottschalk, P., Hinkel, J., ... & Warren, R. F. (2013). A global assessment of the effects of climate policy on the impacts of climate change. *Nature Climate Change*, 3(5), 512 – 519. 2, 3, 4

Balakrishnan, N. (1989). Approximate MLE of the scale parameter of the Rayleigh distribution with censoring. *Reliability, IEEE Transactions on*, 38(3), 355 – 357. 65

Balakrishnan, K. (1996). Exponential distribution: theory, methods and applications. CRC press. 65

Balakrishnan, N., & Aggarwala, R. (2000). Progressive censoring: theory, methods, and applications. Springer. 65

Balasubramanian, K., & Balakrishnan, N. (1992). Estimation for one-and two-parameter exponential distributions under multiple type-II censoring. *Statistical Papers*, 33(1), 203 – 216. 65

Bates, B.C., Kundzewicz, Z.W., Wu, S., Palutikof, J.P. (Eds.) (2008). Climate Change and Water. *IPCC Technical Paper VI*, IPCC, Geneva. 210 pages. 170

Bayarri, M.J., Berger, J.O., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., ... & Walsh, D. (2007). Computer model validation with functional output. *The Annals of Statistics*, 1874 – 1906. 44

Bazzaz, F.A. (1990). The response of natural ecosystems to the rising global CO_2 levels. *Annual Review of Ecological Systematics*, 21, 167 – 196. 22

Bazzaz, F.A., Carlson, R.W. (1984). The response of plants to elevated CO_2: Competition among an assemblage of annuals at two levels of soil moisture. *Oecologia*, 62, 196 – 198. 22

Beniston, M., Stephenson, D. B., Christensen, O. B., Ferro, C. A., Frei, C., Goyette, S., ... & Woth, K. (2007). Future extreme events in European climate: an exploration of regional climate model projections. *Climatic Change*, 81(1), 71 – 95. 3, 36

Benlloch-Gonzalez, M., Bochicchio, R., Berger, J., Bramley, H., & Palta, J. A. (2014). High temperature reduces the positive effect of elevated CO_2 on wheat root system growth. *Field Crops Research*. 16

Bergengren, J.C., Waliser, D.E., Yung, Y.L. (2011). Ecological sensitivity: a biospheric view of climate change. *Climatic Change*, 107(3), 433 – 457. 21, 23, 24, 26

Berridge, C.T., Hadju, L.H., & Dolman, A.J. (2014). How well can we predict soil respiration with climate indicators, now and in the future?. *Biogeosciences Discussions*, 11(2), 1977 – 1999. 104

Berthelot, M., Friedlingstein, P., Ciais, P., Monfray, P., Jufresne, J.L., Le Treut, H., & Fairhead, L. (2002). Global response of the terrestrial biosphere to CO2 and climate change using a coupled climate-carbon cycle model. *Global Biogeochemical Cycles*, 16, 137 – 153. 23, 24, 26

Bondeau, A., Kicklighter, D.W., & Kaduk, K. (1999). Comparing global models of terrestrial net primary productivity (NPP): importance of vegetation

structure on seasonal NPP estimates. *Global Change Biology*, 5(1), 35 – 45. 25

Bondeau, A., Smith, P.C., Zaehle, S., Schaphoff, S., Lucht, W., Cramer, W., Gerten, D., Lotze-Campen, H., Miller, C., Reichstein, M., & Smith, B. (2007). Modelling the role of agriculture for the 20th century global terrestrial carbon balance, *Global Change Biology*, 13(3), 679 – 706. 2, 49

Bornn, L., & Zidek, J.V. (2012). Efficient stabilization of crop yield prediction in the Canadian Prairies. *Agricultural & Forest Meteorology*, 152, 223 – 232. 32, 34, 155

Boukouvalas, A., & Cornfold, D. (2008). Dimension reduction for multivariate emulation. *Technical report*, NCRG/006. 41

Bowman, D.M.J.S., Balch, J.K., Artaxo, P., Bond, W.J., Carlson, J.M., Cochrane, M.A., DAntonio, C.M., DeFries, R.S., et al. (2009). Fire in the Earth System. *Science*, 324(5926), 481 – 484. 14

Box, E.O. (1988). Estimating the seasonal carbon source-sink geography of a mature, steady-state terrestrial biosphere. *Journal of Applied Meteorology*, 27, 1109 – 1124. 11, 12, 91

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of mathematical psychology*, 44(1), 62 – 91. 56

Brailovsky, V. L. (1987). A predictive probabilistic estimate for selecting subsets of regressor variables. *Annals of the New York Academy of Sciences*, 491(1), 233 – 244. 42

Braswell, B.H., & Schimel, D.S., Linder, E., Moore, B. (1997). The response of global terrestrial ecosystems to interannual temperature variability. *Science*, 278, 870 – 872. 30

Bravo, J.I., & De Fuentes, I. (2002). Computing maximum likelihood estimates from Type II doubly censored exponential data. *Statistical Methods and Applications*, 11(2), 187 – 200. 65

Brown, P.J. (1993). Measurememts, Regression, and Calibration. *Oxford Press*, New York. 55, 66

Brown, J.D., Chu, M.T., Ellison, D.C. & Plemmons, R.J. eds. (1994). Proceedings of the Cornelius Lanczos International Centenary Conference (73). *Society*

for Industrial and Applied Mathematics, Philadelphia. Collection; many articles on spectral methods.. 43

Brown, K., & Higginbotham, K.O. (1986). Effects of carbon dioxide enrichment and nitrogen supply on growth of boreal tree seddlings. *Tree Physiology*, 2, 223 – 232. 22

Buckley, J. & James I.A. (1979). Linear regression with censored data. *Biometrika*, 66, 429 – 436. 38, 39

Bunea, F., She, Y., & Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2), 1282 – 1309. 43

Cai, T., & Cheng, S. (2004). Semiparametric regression analysis for doubly censored data. *Biometrika*, 91, 277 – 290. 181

Cao, L., Bala, G., Caldeira, K., Nemani, R., & Ban-Weiss, G. (2009). Climate response to physiological forcing of carbon dioxide simulated by the coupled community atmosphere model (CAM3.1) and community land model (CLM3.0). *Geophysical Research Letter*, 36, L10402. 21

Cao, M., & Woodward, F.I. (1998). Net primary and ecosystem production and carbon stocks of terrestrial ecosystems and their responses to climate change. *Global Change Biology*, 4, 185 – 198. 24, 26

Carter, J.O., Howden, S.M., Day, K.A., & McKeon, G.M. (1998). Soil carbon, nitrogen, phosphorus and biodiversity in relation to climate change. In: Final Report for the Rural Industries *Research and Development Corporation*, 185 – 249. 29, 30

Chang, M. N., & Yang, G. L. (1987). Strong consistency of a nonparametric estimator of the survival function with doubly censored data. *The Annals of Statistics*, 1536 – 1547. 39

Chapin, F.S., Matson, P.A., & Mooney, H.A., (2002). Principles of terrestrial ecosystem ecology. *Springer Science*, New York. 22

Chen, S., & Gopalakrishnan, P. (1998). Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion. *In Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 8. 57

Chen, J., Tian, Y., Zhang, X., Zheng, C., Song, Z., Deng, A., & Zhang, W. (2014). Nighttime warming will increase winter wheat yield through improving plant

development and grain growth in North China. *Journal of Plant Growth Regulation*, 33(2), 397 – 407. 16

Chiles, J.P., & Delfiner, P. (1999). Geostatististics, Modelling Spatial Uncertainty. *Wiley & Sons*, New York. 81

Christensen, J. H., Hewitson, B., Busuioc, A., Chen, A., Gao, X., Held, R., ... & Dethloff, K. (2007). Regional climate projections. *Climate Change, 2007: The Physical Science Basis. Contribution of Working group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, University Press, Cambridge, Chapter 11*, 847-940. 3, 36

Churkina, G., Running, S.W., Schloss, A.L. (1999). Comparing global models of terrestrial net primary productivity (NPP): the importance of water availability. *Global Change Biology*, 5, 46 – 55. 25, 26

Ciais, P., Sabine, C., Bala, G., Bopp, L., Brovkin, V., Canadell, J., Chhabra, A., DeFries, R., Galloway, J., Heimann, M., Jones, C., Le Qur, C., Myneni, R.B., Piao, S., & Thornton, P. (2013). Carbon and Other Biogeochemical Cycles. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. 10

Cleves, M.A., Gould, W.W., Gutierrez, R.G., & Marchenko, Y.U. (2008). An Introduction to Survival Analysis Using Stata. *A Stata Press Publication*, Texas. 123

Cockell, C., Corfield, R., Edwards, R. & Harris, N. (2007). An Introduction to the Earth-Life System. Cambridge University Press. 10

Commenges, D. (2002). Inference for multi-state models from interval censored data. *Statistical Methods*, 11, 167 – 182. 37

Conti, S., Gosling, J. P., Oakley, J. E., & O'Hagan, A. (2009). Gaussian process emulation of dynamic computer codes. *Biometrika*, 96(3), 663 – 676. 45

Conti, S., Gosling, J.P., Oakley, J. & O'Hagan, J. (2008). Gaussian process emulation of dynamic computer codes. *Technical report.* 180

Conti, S., & OHagan, A. (2010). Bayesian emulation of complex multi-output and dynamic computer models. *Journal of Statistical Planning and Inference*, 140(3), 640 – 651. 44, 45

Cook, F.J. & Orchard, V. A. (2008). Relationships between soil respiration and soil moisture, *Soil Biol. Biochem*, 40, 1013 – 1018. 27

Cox, D.R. (1972). Regression models and life-tables (with discussion). *Journal of Royal Statistical Society Series B*, 34, 187 – 202. 39

Cox, P.M., Betts, R.A., Jones, C.D., Spall, S. A., & Totterdell, I.J. (2000). Acceleration of global warming due to carbon cycle feedbacks in a coupled climate model, *Nature*, 408, 184 – 187. 104

Cramer, W., et al. (2001). Global response of terrestrial ecosystem structure and function to CO_2 and climate change: Results from six dynamic global vegetation models, *Global Change Biology*, 7, 357 – 373. 26

Cressie, N. (1993). Statistics for Spatial Data. John Wiley & Sons, New York. 79, 81, 82, 147

Crossette, B., Devi, U., Laski, L., Mahmood, J., Toure, A., & Wong, S. (2010). State of the world population 2010. *United Nations Population Fund*, New York. 31, 186

Currin, C., Mitchell, T., Morris, M., & Ylvisaker, D. (1991). Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association*, 86, 953 – 963. 35, 58

Dancik, G. (2011). mlegp: an R package for Gaussian process modeling and sensitivity analysis. *R package version*, 3(2), 52. 155, 162

David G.T. Denison (Ed.). (2002). Bayesian methods for nonlinear classification and regression (Vol. 386). John Wiley & Sons. 71, 72

Dennis, E. S. & Peacock, W.J. (2009). Vernalization in cereals. *Journal of biology*, 8, 57. 15

De Oliveira, V., Benjamin, K., & David, A. (1997). Bayesian Prediction of Transformed Gaussian Random Fields *Journal of the American Statistical Association*, 92(440), 1422 – 1433. 80

Dillon, G.K., Holden, Z.A., Morgan, P., Crimmins, M.A., Heyerdahl, E.K., & Luce, C.H., (2011). Both Topography and Climate Affected Forest and Woodland Burn Severity in Two Regions of the Western US, 1984 to 2006. *Ecosphere*, 2(12), art130. 14

Dobermann, A. & Nelson, R. (2013). Opportunities and solutions for sustainable food production. *Sustainable Development Solutions Network*, Paris, France. 31

Draper, N.R. & Smith, H. (1980). Applied Regression Analysis. *John Wiley & Sons*, USA. 55, 58

Drineas, P. & Mahoney, M.W. (2005a). On the Nystrom Method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6, 2153 – 2175. 155

Drineas, P., & Mahoney, M. W. (2005b). Approximating a gram matrix for improved kernel-based learning. *In Learning Theory (*323−337*)*. Springer Berlin Heidelberg. 43

Drummond, S.T., Sudduth, K.A., Joshi, A., Birrell, S.J. & Kitchen, N.R. (2003). Statistical and neural methods for site-specific yield prediction. *American Society of Agricultural Engineers*, 46(1). 32, 34

Dubey, S.K. (2001). Methane emission and rice agriculture. *Current Science*, 81, 345 – 346. 17

Dufresne, J.L., Friedlingstein, P., Berthelot, M., Bopp, L., Ciais, P., L. Fairhead, L., & Monfray, P. (2002). On the magnitude of positive feedback between future climate change and the carbon cycle, *Geophysics Research Letter*, 29, 1405. 23, 26, 58

Ecker, M. D. & Gelfand, A. E. (1997). Bayesian variogram modeling for an isotropic spatial process. *Journal of the Agricultural, Biological & Environmental Statistics*, 2, 347 – 369. 82

Fader, M., Rost, S., Mller, C., Bondeau, A. and Gerten, D. (2010). Virtual water content of temperate cereals and maize: Present and potential future patterns. *Journal of Hydrology*, 384, 218 – 231. 2, 49, 51, 52, 53

Faraway, J.J. (2006). Extending the linear model with R: generalized linear, mixed effects, and nonparametric regression models. *Chapman & Hall/CRC*, Boca Raton, Florida. 35, 55, 66

Field, C.B. et al. (1998). Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science*, 281, 237 – 240. 59

Finley, A. O., Banerjee, S., & Basso, B. (2011). Improving crop model inference through Bayesian melding with spatially varying parameters. *Journal of Agricultural, Biological, and Environmental Statistics*, 16(4), 453 – 474. 155

Fischer, G., Shah, M., van Velthuizen, H., & Nachtergaele, F. O. (2001). Global agro-ecological assessment for agriculture in the 21st century. IIASA Research Report 02-02, International Institute for Applied Systems Analysis, Laxenburg, Austria, p. 119. 2

Floresa, E.I., DiazDelaO, F.A., Friswell, M.I., & Ajaj, R.M. (2014). Investigation on the extensibility of the wood cell-wall composite by an approach based on homogenisation and uncertainty analysis. *Composite Structures*, 212 – 222. 77

Foster, D. & George, E. (1993). The risk inflation factor in multiple linear regression. *Technical Report, Univ. of Texas.* 57

Foster, D. P., & George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, 1947 – 1975. 56

Fricker, T., Oakley, J., & Urban, N. M. (2010). Multivariate emulators with nonseparable covariance structures. URL http://www. mucm. ac. uk/Pages/Dissemination/Dissemination- Papers-Technical. html. 44, 45

Frydman, H., & Liu, J. (2013). Nonparametric estimation of the cumulative intensities in an interval censored competing risks model. *Lifetime data analysis*, 19(1), 79 – 99. 39

Gerten, D., Schaphoff, S., Haberlandt, U., Lucht, W., & Sitch, S. (2004). Terrestrial vegetation and water balance hydrological evaluation of a dynamic global vegetation model. *Journal of Hydrology*, 286, 249 – 270. 49

Giardina, C.P. & Ryan, M.G. (2000). Evidence that decomposition rates of organic carbon in mineral soil do not vary with temperature. *Nature*, 404(6780), 858 – 861. 30

Gibbs, M., & MacKay, D.J. (1997). Efficient implementation of Gaussian processes. *Technical Report.* http://citeseerx.ist.psu.edu/viewdoc/summary. 43

Godfray, H.C.J., Beddington, J.R., Crute, I.R., Haddad, L. & Lawrence, D. et al. (2010). Food security: the challenge of feeding 9 billion people. *Science*, 327, 812 – 818. 31

Goldstein, M. (1995). Bayes linear methods I-Adjusting beliefs: concepts and properties. unpublished tutorial http://maths. dur. ac. uk/stats/bd. 73

Goldstein, M., & Rougier, J. (2006). Bayes linear calibrated prediction for complex systems. *Journal of the American Statistical Association*, 101(475). 73

Gopalakrishnan, S., Kumar, R. M., Humayun, P., Srinivas, V., Kumari, B. R., Vijayabharathi, R., & Rupela, O. (2014). Assessment of different methods of rice (Oryza sativa. L) cultivation affecting growth parameters, soil chemical, biological, and microbiological properties, water saving, and grain yield in ricerice system. *Paddy and Water Environment*, 12(1), 79 – 87. 17

Gosling, J.P. & O'Hagan, A. (2007). Understanding the uncertainty in the biospheric carbon flux for England and Wales. *Technical Report*, 567/06, Department of Probability and Statistics, University of Sheffield. 3, 40, 58

Goward, S.N., Masek, J.G., Cohen, W., Moisen, G., et al. (2008). Forest disturbance and North American carbon flux. *Eos, Transactions American Geophysical Union*, 89(11), 105 – 106. 14

Graham, L. P., Hagemann, S., Jaun, S., & Beniston, M. (2007). On interpreting hydrological change from regional climate models. *Climatic Change*, 81(1), 97 – 122. 3, 36

Handcock, M. S. & Stein, M. L. (1993). Bayesian analysis of kriging. *Technometrics*, 35, 403 – 410. 80, 81

Hankin, R. K. (2012). Introducing multivator: A Multivariate Emulator. *Journal of Statistical Software*, 46(8), 1 – 20. 45

Hannah, L. A., Blei, D. M., & Powell, W. B. (2011). Dirichlet process mixtures of generalized linear models. *The Journal of Machine Learning Research*, 12, 1923 – 1953. 42

Heitmann, K., Higdon, D., Nakhleh, C., & Habib, S., (2006). Cosmic Calibration. *The Astrophysical Journal*, 646, 2, $L1 - L4$. 44, 155

Heyder, U., Schaphoff, S., Gerten, D., & Lucht, W. (2011). Risk of severe climate change impact on the terrestrial bisosphere. *Environmental Research Letter*, 6, 034036. 3, 21, 26

Higdon, D.M., Gattiker, J., Williams, B. & Rightley, M. (2008). Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, 103, 570 – 583. 41, 44, 155

Higdon, D.M., Lee, H. & Holloman, C. (2003). Markov chain Monte Carlo-based approaches for inference in computationally intensive inverse problems, in Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F., and West, M. (eds), Bayesian Statistics. *Proceedings of the Seventh Valencia International Meeting*, Oxford University Press, 181 – 197. 35

Hirano, T. (2014). Modified Linear Projection for Large Spatial Data Sets. *arXiv preprint arXiv:1402.5847.* 43

Holden, P.B., Edwards, N.R., Oliver, K.I.C., Lenton, T.M. & Wilkinson, R.D. (2010a). A probabilistic calibration of climate sensitivity and terrestrial carbon change in GENIE-1, Climate Dynamics, 35, 785 – 806. 3, 91

Holden, P.B., & Edwards, N.R. (2010b). Dimensionally reduced emulation of an AOGCM for application to Integrated Assessment Modelling, *Geophysical Research Letters*, 37. 3, 91

Holden, P.B., & Edwards, N.R., Garthwaite, P.H., Fraedrich, K., Lunkeit, F., Kirk, E., ... & Babonneau, F. (2014). PLASIM-ENTSem v1. 0: a spatio-temporal emulator of future climate change for impacts assessment. *Geoscientific Model Development*, 7(1), 433 – 451. 109

Holland, E.A., Neff, J.C., Townsend, A.R., & McKeown, B. (2000) Uncertainties in the temperature sensitivity of decomposition in tropical and subtropical ecosystems: Implications for models. *Global Biogeochemical Cycles*, 14(4), 1137 – 1151. 29

Horion, S., Cornet, Y., Erpicum, M. & Tychon, B. (2013). Studying interactions between climate variability and vegetation dynamic using a phenology based approach. *International Journal of Applied Earth Observation and Geoinformation*, 20, 20 – 32. 28, 33

Iglesias, A., Rosenzweig, C., & Pereira, D., (2000). Prediction spatial impacts of climate in agriculture in Spain. *Global Environmental Change*, 10, 69 – 80. 2

Imhoff, M.L., Bounoua, L., & DeFries, R.S. et al. (2004) The consequences of urban land transformations on net primary productivity in the United States. *Remote Sensing of the Environment*, 89, 434 – 443. 12, 26

IPCC WGI Fourth Assessment Report (2007): Climate Change: *The Physical Science Basis*, in Summary for Policymakers by Alley, R., Berntsen, T., Nathaniel, R., Bindoff, L. et al. 1, 10

Narciso, J., & Hossain, M. (2002). World rice statistics. Available by IRRI www. irri. org/science/ricestat (posted2002). 17

Ise, T., & Moorcroft, P.R. (2006). The global-scale temperature and moisture dependencies of soil organic carbon decomposition: an analysis using a mechanistic decomposition model. *Biogeochemistry*, 80, 217 – 231. 27

Jackson, C.S., Sen, M., & Stoffa, P. (2004). An efficient stochastic Bayesian approach to optimal parameter and uncertainty estimation for climate model predictions. *Journal of Climate*, 17, 2828 – 2840. 35

Jain, R.C. & Agrawal, R. (1992). Probability model for crop yield forecasting. *Biometrical Journal*, 34(4), 501 – 511. 33

Jammalamadaka, S. R., & Mangalam, V. (2003). Nonparametric estimation for middle-censored data. *Journal of Nonparametric Statistics*, 15(2), 253 – 265. 39

Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A., Wilkens, P.W., Singh, U., Gijsman, A.J., & Ritchie, J.T., (2003). The DSSAT cropping system model. *European Jounal of Agronomy.* 18, 235–265. 2

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457 – 481. 38

Kaplan, E. L., & Meier, P. (1992). Nonparametric estimation from incomplete observations. *In Breakthrough in Statistics*, 319 – 337. Springer New York. 38

Kart, R.W. (1979). Sensitivity analysis of statistical crop weather models. *Agricultural Meteorology*, 20, 291 – 300. 32

Katterer, T., Reichstein, M., Andren, O. & Lomander, A. (1998). Temperature dependence of organic matter decomposition: a critical review using literature data analyzed with different models. *Biology and Fertillty of Soils*, 27, 258 – 262. 29

Keeling, C.D., Chin, J.F.S., & Whorf, T.P. (1996). Increased activity of northern vegetation inferred from atmospheric CO_2 measurements. *Nature*, 382(6587), 146 – 149. 11

Kelley, D.I., Harrison, S.P., & Prentice, I. C. (2014). Improved simulation of fire-vegetation interactions in the Land surface Processes and eXchanges dynamic global vegetation model (LPX-Mv1). *Geoscientific Model Development Discussions*, 7(1), 931 – 1000. 14

Kennedy, M.C. & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of Royal Statistical Society*, series B, 63(3), 425 – 464. 10, 22, 35, 41

Kennedy, M.C., Anderson, C. W., Conti, S., & O'Hagan, A. (2006). Case studies in Gaussian process modelling of computer codes. *Reliability Engineering & System Safety*, 91, 1301 – 1309. 42

Kennedy, M.C. et al. (2008). Quantifying uncertainty in the biospheric carbon flux for England and Wales. *Journal of the Royal Statistical Society*, Series B, 171, 109 – 135. 40

Kenneth Y.C., & James L.P. (2001). Semiparametric censored regression models *Journal of Economic Perspectives*, 15(4), 29 – 42. 123

Kettleborough, J., Booth, B., Stott, P., & Allen, M. (2007). Estimates of uncertainty in predictions of global mean surface temperature. *Journal of Climate*, 20, 843 – 855. 35, 36

Kim, Y., Lee, J., & Kim, J. (2005). Bayesian bootstrap analysis of doubly censored data using Gibbs sampler. *Statistica Sinica*, 969 – 980. 33, 39

Kim, D. H., Doyle, M. R., Sung, S., & Amasino, R. M. (2009). Vernalization: winter and the timing of flowering in plants. *Annual Review of Cell and Developmental*, 25, 277 – 299. 15

King, A.W., Post, W.M., & Wullschleger, S.D. (1997). The potential response of terrestrial carbon storage to changes in climate and atmospheric CO_2, *Climate Change*, 35, 199 – 227. 21, 26

Kirschbaum, M.U.F. (2000) Will changes in soil organic carbon act as a positive or negative feedback on global warming? *Biogeochemistry*, 48, 21 – 51. 21, 30

Kirschbaum, M.U.F., Eamus, D., Gifford, R.M., Roxburgh, S.H., & Sands, P.J. (2001). Definitions of some ecological terms commonly used in carbon

accounting. *Cooperative Research Centre for Carbon Accounting*, Canberra, 2 – 5. 13

Kitchen, N.R., Sudduth,K.A., & Drummond, S.T. (1999.) Electrical conductivity as a crop productivity measure for claypan soils. *Journal of Prod. Agriculture*, 12(4), 607617. 33

Knutti, R., & Sedlacek, J. (2013). Robustness and uncertainties in the new CMIP5 climate model projections. *Nature Climate Change*, 3(4), 369 – 373. 185

Krantgartner, R., Nenard, M. C., Lieberz, S., Boshnakova, M., Flach, B., Wilson, J., Wideback, A., Bettini, O., Guerrero, M., & Bendz, K. (2011): EU-27 Oil seeds and products Annual modest rebound in EU-27, Oils seeds Products, GAIN Report No. E 60016, USDA. 18

Lai, T. L., & Ying, Z. (1991). Large sample theory of a modified Buckley-James estimator for regression analysis with censored data. *The Annals of Statistics*, 1370 – 1402. 38

Lampe, K. (1995). Rice research: food for 4 billion people. *Geography Journal*,35, 253 – 259. 16

Lee, K. R., Kapadia, C. H., & Brock, D. B. (1980). On estimating the scale parameter of the Rayleigh distribution from doubly censored samples. *Statistische Hefte*, 21(1), 14 – 29. 65

Lee, L. A., Carslaw, K. S., Pringle, K. J., & Mann, G. W. (2012). Mapping the uncertainty in global CCN using emulation. *Atmospheric Chemistry and Physics*, 12(20), 9739 – 9751. 157

Lee, L. A., Pringle, K. J., Reddington, C. L., Mann, G. W., Stier, P., Spracklen, D. V., & Carslaw, K. S. (2013). The magnitude and causes of uncertainty in global model simulations of cloud condensation nuclei. *Atmospheric Chemistry & Physics Discussions*, 13(3). 157

Leemans, R., & Solomon, A. (1993), Modeling the potential change in yield and distribution of the earth's crops under a warmed climate. *Climate Research*, 3, 79 – 96. 2

Le Quere, C., Peters, G. P., Andres, R. J., Andrew, R. M., Boden, T., Ciais, P., & Harper, A. (2013). Global carbon budget 2013. *Earth System Science Data Discussions*, 6(2). 10

Lieth, H.F.H. (1975). Primary production of the major vegetation units of the world. In: *Primary Productivity of the Biosphere* (eds Lieth, H., Whittaker, R.H.), 14, 237 – 263. *Springer-Verlag*, New York. 26

Lin, G., He, X., & Portnoy, S. (2012). Quantile regression with doubly censored data. *Computational Statistics & Data Analysis*, 56(4), 797 – 812. 40

Liu, S., Yang, J.Y., Zhang, X.Y., Drury, C.F. and Reynolds, W.D. et al. (2013). Modelling crop yield, soil water content and soil temperature for a soybean-maize rotation under conventional and conservation tillage systems in Northeast China. *Agricultural Water Management*, 123, 32 – 44. 33, 34

Lloyd, J., & Taylor, J.A. (1994). On the temperature dependence of soil respiration. *Functional Ecology*, 8, 315 – 323. 27, 28, 30, 104

Lobell, D.B. (2013). Errors in climate datasets and their effects on statistical crop models. *Agricultural and Forest Meteorology*, 170, 58 – 66. 33

Lobell, D.B. & Burke, M.B. (2008). Why are agricultural impacts of climate change so uncertain? The importance of temperature relative to precipitation. *Environmental Research Letter*, 3, 034007. 3, 153, 154, 182

Lobell, D.B., & Burke, M.B. (2010). On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology*, 150(11), 1443 – 1452. 3, 154, 182

Lobell, D.B., Cahill, K.N, & Field, C.B. (2006). Weather-based yield forecasts developed for 12 California crops *California Agriculture*, 60(4), 210 – 214. 33

Lobell, D. B., & Field, C.B. (2007). Global scale climate-crop yield relationships and the impacts of recent warming.*Environmental Research Letter*, 2, 1 – 7. 3

Loehman, R. A., Reinhardt, E., & Riley, K. L. (2013). Wildland fire emissions, carbon, and climate: Seeing the forest and the treesA cross-scale assessment of wildfire and carbon dynamics in fire-prone, forested ecosystems. *Forest Ecology and Management*, 317, 9 – 19. 14

Lopez, A., Tebaldi, C., New, M., Stainforth, D., Allen, M., & Kettleborough, J. (2006). Two approaches to quantifying uncertainty in global temperature changes. *Journal of Climate*, 19, 4785 – 4796. 40

Biegler, L., Biros, G., Ghattas, O., Heinkenschloss, M., Keyes, D., Mallick, B., Luis, T. et al. (Eds.). (2011). Large-scale Inverse Problems and Quantification of Uncertainty. *John Wiley & Sons*, UK. 5

Luo, Z., & Wahba, G. (1997). Hybrid adaptive splines. *Journal of the American Statistical Association*, 92(437), 107 – 116. 42

Mahecha, M. D., Reichstein, M., Carvalhais, N., Lasslop, G., Lange, H., Seneviratne, S. I., & Richardson, A. D. (2010). Global convergence in the temperature sensitivity of respiration at ecosystem level. *Science*, 329(5993), 838 – 840. 27, 28, 31

Matis, J.H., Birkett, T. & Boudreaux, D. (1989). An application of the Markov chain approach to forecasting cotton yields from Surveys. *Agricultural Systems*, 29, 357 – 370. 33

Matthews, E., Fung, I., & Lerner, J. (1991). Methane emission from rice cultivation: Geographic and seasonal distribution of cultivated areas and emissions. *Global Biogeochemical Cycles*, 5(1), 3 – 24. 17

Maxwell, J.T., Knapp, P.A. & Ortegren, J.T. (2013). Influence of the Atlantic Multidecadal Oscillation on tupelo honey production from AD 1800 to 2010. *Agricultural and Forest Meteorology*, 174 – 175, 129 – 134. 27, 33

McGuire, A.D., Joyce, L.A., Kicklighter, D.W., Melillo, J.M., Esser, G., & Vorosmarty, C.J., (1993). Productivity response of climax temperate forests to elevated temperature and carbon dioxide: a North American comparison between two global models. *Climate Change*, 24, 287 – 310. 11

McGuire, A.D., Sitch, S., Clein J.S., et al. (2001). Carbon balance of the terrestrial biosphere in the twentieth century: analyses of CO_2, climate and land use effects with four process-based ecosystem models. *Global Biogeochemical Cycles*, 15(1), 183 – 206. 24

Meinshausen, M., Smith, S. J., Calvin, K., Daniel, J. S., Kainuma, M.L.T., Lamarque, J. F., ... & Van Vuuren, D. P. P. (2011a). The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. *Climatic change*, 109(1 – 2), 213 – 241. 19

Meinshausen, M., Raper, S.C., & Wigley, T.M. (2011b). Emulating coupled atmosphere-ocean and carbon cycle models with a simpler model, MAGICC6 Part 1: Model description and calibration. *Atmospheric, Chemistry and Physics*, 11, 1417 – 1456. 48

Melillo, J.M., McGuire, A.D., et al. (1993). Global climate change and terrestrial net primary production. *Nature*, 363, 234 – 240. 11, 22, 24, 25, 26

Miller, R.G. (1976). Least squares regression with censored data. *Biometrika*, 63(3), 449 – 464. 32, 38, 39

Miller, R., & Halpern, J. (1982). Regression with censored data. *Biometrika*, 69(3), 521 – 531. 38, 39

Milner, K.S., Running, S.W., & Coble, D.W. (1996). A biophysical soil-site model for estimating potential productivity of forested landscapes. *Canadian Journal of Forest Resources*, 26, 1174 – 1186. 11

Mitchell, T.D., Carter, T.R., Jones, P.D., Hulme, M. & M. New (2004). A comprehensive set of high-resolution grids of monthly climate for Europe and the globe: The observed record (1901-2000) and 16 scenarios (2001-2100). *Tyndall Centre Working Paper 55*, Tyndall Centre for Climate Change Research, University of East Anglia, Norwich, 30pp. 48

Mitchell, R. J., Liu, Y., OBrien, J. J., Elliott, K. J., Starr, G., Miniat, C. F., & Hiers, J. K. (2014). Future climate and fire interactions in the southeastern region of the United States. *Forest Ecology and Management*, In Press. 14

Monod, H., Naud, C., & Makowski, D. (2006). Uncertainty and sensitivity analysis for crop models. D. Wallach, D. Makowski et J. Jones, eds: *Working with Dynamic Crop Models*, 55 – 100. 10, 83

Moorcroft, P.R. (2006) How close are we to predictive science of the biosphere? *Trends in Ecology and Evolution*, 21(7), 401 – 407. 23, 24, 26

Moore, M.R., Gollehon, N.R. & Hellerstein, D.M. (2000). Estimating producer's surplus with the censored regression model: An application to producers affected by Columbia River Basin Salmon Recovery. *Journal ofAgricultural and Resource Economics*, 25(2), 325 – 346. 181

Morris, J. S., Vannucci, M., Brown, P. J. & Carroll, R. J. (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis. *Journal of the American Statistical Association*, 98, 573 – 583. 44

Morris, J. S., & Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2), 179 – 199. 44

Moss, R.H., Edmonds, J.A., Hibbard, K.A., Manning, M.R. and Rose, S.K. et al. (2010). The next generation of scenarios for climate change research and assessment. *Nature*, 463, 747 – 756. 20

Moyano, F. E., Vasilyeva, N., Bouckaert, L., Cook, F., Craine, J., Curiel Yuste, J., & Chenu, C. (2012). The moisture response of soil heterotrophic respiration: interaction with soil properties. *Biogeosciences*, 9(3), 1173 – 1182. 27

Muller, C. & Robertson, R. (2014): Projecting future crop productivity for global economic modeling. *Agricultural Economics*, 45, 1, 37 – 50 2

Murphy, K. P. (2007). Conjugate Bayesian analysis of the Gaussian distribution. *Technical Report*, University of British Columbia. 71, 72, 77

Nakicenovic, N., & Swart, R. (2000). Special report on emissions scenarios. Edited by Nebojsa Nakicenovic and Robert Swart, pp. 612. ISBN 0521804930. Cambridge, UK: Cambridge University Press, July 2000., 1. 19

Neilson, R.P., Marks, D. (1994). A global perspective of regional vegetation and hydrologic sensitivities from climate change. *Journal of Vegetation Science*, 5, 715 – 730. 26

Nelson, G. C., Valin, H., Sands, R. D., Havlik, P., Ahammad, H., Deryng, D., ... & Willenbockel, D. (2014). Climate change effects on agriculture: Economic responses to biophysical shocks. *Proceedings of the National Academy of Sciences*, 111(9), 3274 – 3279. 2

Nemani, R.R., Keeling, C.D., Hashimoto, H., Jolly, W.M., Piper, S.C., Tucker, C.J., Myneni, R.B., & Running, S.W. (2003). Climate-driven increases in global terrestrial net primary production from 1982-1999. *Science Report*, 300. 26

Oakley, J. (1999). *Bayesian Uncertainty Analysis For Complex Computer Codes.* Ph.D. thesis, University of Sheffield. 45, 77

Oakley, J. and O'Hagan, A. (2002). Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika*, 89, 769 – 784. 35, 41

Oakley, J. E. and O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Journal of Royal Statistical Society*, 66B, 751– 769. 41, 43, 44, 180

O'Hagan, A. (2006). Bayesian Analysis of Computer Code Outputs: A Tutorial. *Reliability Engineering and System Safety*, 91, 1290 – 1300. 8, 10, 105, 180

Oliver, S. N., Finnegan, E. J., Dennis, E. S., Peacock, W. J., & Trevaskis, B. (2009). Vernalization-induced flowering in cereals is associated with changes in histone methylation at the *Vernalization1* gene. *Proceedings of the National Academy of Sciences*, 106(20), 386 – 8391. 15

Osborn, T.J. (2009). A user guide for ClimGen: a flexible tool for generating monthly climate data sets and scenarios. *Climatic Research Unit*, University of East Anglia, Norwich, 17. 48

Osborne, T.M., Rose, G. & Wheeler, T.R. (2013). Variation in global-scale impacts of climate change on crop productivity due to climate model uncertainty and adaptation. *Agriculutural & Forest Meteorology*, 170, 183 – 194. 33, 34, 36, 153

Osborne, T.M. & Wheeler, T.R. (2013). Evidence for a climate signal in trends of global crop yield variability over the past 50 years. *Environmental Research Letter*,8, 024001. 27, 33

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, 57(1), 120 – 125. 57

Parry, M.L., Fischer, C., Livermore, M., Rosenzweig, C., & Iglesias, A. (1999). Climate change and world food security: a new assessment. *Global Environmental Change*, 9, $S51 - S67$. 2

Parry, M.L., Rosenzweig, C., Iglesias, A., Livermore, M. & Fischer, G. (2004). Effects of climate change on global food production under SRES emissions and socioe conomic scenarios. *Global Environmental Change*, 14, 53 – 67. 32, 34, 186

Parry, M., Rosenzweig, C., & Livermore, M. (2005). Climate change, global food supply and risk of hunger. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1463), 2125 – 2138. 2, 170

Parry, M. L. (Ed.). (2007). Climate Change 2007: impacts, adaptation and vulnerability: contribution of Working Group II to the fourth assessment report of the Intergovernmental Panel on Climate Change (Vol. 4). Cambridge University Press. 170

Passioura, J. B., & Angus, J. F. (2010). Improving productivity of crops in water-limited environments. *Advances in agronomy*, 106, 37 – 75. 16

Peng, C.H., Guiot, J., Van Campo, E., & Cheddadi, R., (1995). Temporal and spatial variations of terrestrial biomes and carbon storage since 13 000 years BP in Europe: reconstruction from pollen data and statistical models. *Water Air Soil Pollution*, 82, 375 – 391. 11

Peng, C., & Apps, M.J. (1999). Modelling the response of net primary productivity (NPP) of boreal forest ecosystems to changes in climate and fire disturbance regimes. *Ecological Modelling*, 122, 175 – 193. 25, 26

Poggio, T. & Girosi, F. (1990). Networks for Approximation and Learning. *Proceedings of IEEE*, 78, 1481 – 1497. 42

Porter, J. R., & Gawith, M. (1999). Temperatures and the growth and development of wheat: a review. *European Journal of Agronomy*, 10(1), 23 – 36. 15

Posada, D., & Buckley, T R. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53(5), 793 – 808. 57

Potter, C.S. (1993). Terrestrial ecosystem production: a process model based on global satellite and surface data. *Global Biogeochemical Cycles*, 7(4), 811 – 841. 24

Potter, U. (2000). A multivariate Buckley-James estimator, In: Kollo, T., Tiit, E.M., Srivastava, M. (Eds.), Multivariate Statistics. VSP, 117131. 38

Prince, S.D. (1991). A model of regional primary production for use with coarse-resolution satellite data. *International Journal of Remote Sensing*, 12, 1313 – 1330. 25, 26

Quinonero-Candela, J., & Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, 6, 1939 – 1959. 42, 72

R: A Language and Environment for Statistical Computing, R Core Team,R Foundation for Statistical Computing, Vienna, Austria, 2013, ISBN 3-900051-07-0, http://www.R-project.org/. 91, 92, 124, 150

Raich, J.W., & Schlesinger, W.H. (1992). The global carbon dioxide flux in soil respiration and its relationship to vegetation and climate. *Tellus B*, 44(2), 81 – 99. 13

Raich, J. W., Potter, C. S., & Bhagawati, D. (2002). Interannual variability in global soil respiration, 1980-94. *Global Change Biology*, 8(8), 800 – 812. 104

Ramesh, K. V., & Goswami, P. (2014). Assessing reliability of regional climate projections: the case of Indian monsoon. *Scientific reports*, 4. 185

Rasmussen, C.E. (1999). The infinite Gaussian mixture model. *Advances in Neural Information Processing Systems*, 12, 554 – 560. 42

Rasmussen, C.E. and Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning*, the MIT Press. 42, 72, 73, 78, 155

Ream, T.S., Woods, D.P., Schwartz, C.J., Sanabria, C. P., Mahoy, J. A., Walters, E. M., & Amasino, R. M. (2014). Interaction of photoperiod and vernalization determines flowering time of Brachypodium distachyon. *Plant physiology*, 164(2), 694 – 709. 15

Reddy, V.R. & Pachepsky, Y.A. (2000). Predicting crop yields under climate change conditions from monthly GCM weather projections. *Environmental Modelling & Software*, 15, 79 – 86. 32, 34

Reich, J.W. & Schlesinger, W.H. (1992). The global carbon dioxide flux in soil respiration and its relationship to climate. *Tellus B*, 44(2), 81 – 996. 29, 30

Ren, J. & Gu, M. (1997). Regression M-estimators with doubly censored data. *The Annals of Statistics*, 25, 2638 – 2664. 39

Ren, J.J. (2003). Regression M-estimators with non-iid doubly censored data. *The Annals of Statistics*, 31(4), 1186 – 1219. 40

Ren, J.J. (2008). Weighted empirical likelihood in some two-sample semiparametric models with various types of censored data. *The Annals of Statistics*, 36, 147 – 166. 40

Ricker, J.B. (2011). The RevoScaleR Data Step. *White Paper.* http://www.revolutionanalytics.com/sites/default/files/data-step-white-paper.pdf. 28

Riedel, K. S. (1992). A Sherman-Morrison-Woodbury identity for rank augmenting matrices with application to centering. *SIAM Journal on Matrix Analysis and Applications*, 13(2), 659 – 662. 43

Ritov, Y. A (1990). Estimation in a linear regression model with censored data. *The Annals of Statistics*, 18, 303 – 328. 38, 39

Rosenzweig, M.L. (1968). Net primary productivity of terrestrial communities: prediction from climatological data. *The American Naturalist*, 102, 67 – 74. 26

Rosenzweig, C., & Parry M.L. (1994). Potential impact of climate change on world food supply. *Nature*, 367, 133 – 138. 2, 170

Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Mller, C., Arneth, A., ... & Jones, J. W. (2014). Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. *Proceedings of the National Academy of Sciences*, 111(9), 3268 – 3273. 2

Rougier, J.C. (2007). Probabilistic inference for future climate using an ensemble of climate model evaluations. *Climate Change*, 81, 247 – 264. 40

Rougier, J. (2008). Efficient emulators for multivariate deterministic functions. *Journal of Computational and Graphical Statistics*, 17(4), 827 – 843. 45, 155

Rougier, J., Guillas, S., Maute, A., & Richmond, A. D. (2009). Expert knowledge and multivariate emulation: The thermosphereionosphere electrodynamics general circulation model (TIE-GCM). *Technometrics*, 51(4), 414 – 424. 45, 155

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational and Behavioral Statistics*, 6(4), 377 – 401. 39

Ryan, M.G. (1991). Effects of climate change on plant respiration. it Ecological Applications, 1(2), 157 – 167. 29

Ryan, M.G., Lavigne, M.B., & Gower, S.T. (1997). Annual carbon cost of autotrophic respiration in boreal forest ecosystems in relation to species and climate. *Journal of Geophysical Research: Atmospheres*, 102(D24), 28871 – 28883. 29

Sacks, J., Welch, W., Mitchell, T., & Wynn, H. (1998). Design and analysis of computer experiments. *Statistical Science*, 4(4), 409 – 435. 3, 10, 72, 79

Sacks,W.J., Deryng, D., Foley, J.A., & Ramankutty, N. (2010). Crop planting dates: an analysis of global patterns. *Global Ecology and Biogeography*, 19, 607620. 3, 33, 34

Saltelli, A. (2002), Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communication*, 145, 580 – 297. 84, 92

Saltelli, A., Chan, K., & Scott, E. M. (Eds.). (2000). Sensitivity analysis (Vol. 134). New York: Wiley. 83, 86

Sang, H., & Huang, J. Z. (2012). A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society*, Series B (Statistical Methodology), 74(1), 111 – 132. 43

Santner, T., Williams, B., & Notz, W. (2003). *The Design and Analysis of Computer Experiments.* Springer. 35, 72, 78, 84, 92

Schabenberger, O., & Gotway, C.A. (2005). *Statistical Methods for Spatial Data Analysis.* Boca Raton: Chapman & Hall/CRC. 81

Schindlbacher A., Zechmeister-Boltenstern S., & Jandl, R. (2009). Carbon losses due to soil warming: do autotrophic and heterotrophic soil respiration respond equally? *Global Change Biology*, 15, 901 – 913. 29

Schlenker, W., Michael J. & Roberts, M.J. (2006). Estimating the impact of climate change on crop yields: The importance of non-linear temperature effects. *Discussion Paper*, 0607-01, Department of Economics Columbia University New York, NY. 32, 33, 34

Schlenker, W., & Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to US crop yields under climate change. *Proceedings of the National Academy of Sciences*, 106(37), 15594 – 15598. 3

Schloss, A.L., Kicklighter, D.W., Kaduk, K., & Wittenberg, U. (1999). Comparing global models of terrestrial net primary productivity (NPP): comparison of NPP to climate and Normalized Difference Vegetation Index (NDVI). *Global Change Biology*, 5, 25 – 34. 25

Schnedler, W. (2005). Likelihood estimation for censored random vectors. *Econometric Reviews*, 24(2), 195 – 217. 37, 38

Schwarz, G. (1978). Estimating the dimension of a model. *Annal of Statistics.* 6, 461 – 464. 57

Seeger, M. (2000). Skilling techniques for Bayesian analysis. *Technical Notes*, University of Edinburgh. 43

Shahsavani, D. & Grimvall, A. (2011). Variance-based sensitivity analysis of model outputs using surrogate models. *Environmental Modelling & Software*, 26, 723 – 730. 35

Shakun, J.D., Clark, P.U., Marcott, S.A. et al. (2012). Global warming preceded by increasing carbon dioxide concentrations during the last deglaciation. *Nature*, 484, 49 – 54. 21

Shen, P.S. (2013). Regression M-estimators with one modified form of doubly censored data. *Communications in Statistics-Simulation and Computation*, 42(3), 603 – 612. 40

Shi, J. Q., Murray-Smith, R., & Titterington, D.M. (2003). Bayesian regression and classification using mixtures of Gaussian processes. *International Journal of Adaptive Control and Signal Processing*, 17(2), 149 – 161. 42

Shi, J.Q., Murray-Smith, R. and Titterington, D.M. (2005). Hierarchical Gaussian Process mixtures for regression. *Statistics and Computing*, 15, 31 – 41. 42

Shi, J.Q., Murray-Smith, R., Titterington, D.M., & Pearlmutter, B. A. (2005). Filtered gaussian processes for learning with large data-sets. *In Switching and Learning in Feedback Systems*, 128 – 139. Springer Berlin Heidelberg. 43, 45

Shrestha, S., Deb, P., & Bui, T. T. T. (2014). Adaptation strategies for rice cultivation under climate change in Central Vietnam. *Mitigation and Adaptation Strategies for Global Change*, 1 – 23. 17

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *Journal of Royal Statistical Society, series B*,47(1), 1 – 52. 155

Sitch, S., Smith, B., Prentice I.C., Arneth, A. and Bondeau, A. et al. (2003). Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. *Global Change Biology*, 9, 161 – 185. 49

Skilling, J. (1993). Bayesian numerical analysis. *Physics and probability, Cambridge*, CUP. 43, 45

Smith P. & Gregory P.J. (2013). Climate change and sustainable food production. *The Proceedings of the Nutrition Society*, 72(1), 21 – 28. 27, 28, 31

Sobol, I.M. (1993). Sensitivity analysis of non-linear mathematical models. *Mathathematical Modelling & Computational Experiment*, 1, 407 – 414. 84, 85

Solin, A., & Sarkka, S. (2014). Hilbert Space Methods for Reduced-Rank Gaussian Process Regression. Department of Biomedical Engineering and Computational Science (BECS). *arXiv preprint arXiv:1401.5508*. 43

Solomon, S. (2007) (ed) Climate Change: The Physical Science Basis. *Contribution of Working Group I to the Fourth Assessment Report of the Intergovermental Panel on Climate Change*. Cambridge University Press, UK. 48

Sommers, W. T., Loehman, R. A., & Hardy, C. C. (2014). Wildland fire emissions, carbon, and climate: Science overview and knowledge needs. *Forest Ecology and Management*, 317, 1 – 8. 14

South, A. (2011). rworldmap: a new R package for mapping global data. *The R Journal*, 3(1), 35 – 43. 163

Steduto, P., Hsiao, T. C., Fereres, E., & Raes, D. (2012). Crop yield response to water. *FAO Irrigation and Drainage Paper*, 66. 15

Stocker, T. F., Dahe, Q., & Plattner, G. K. (2013). Climate Change 2013: The Physical Science Basis. Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Summary for Policymakers (IPCC, 2013). 20, 22, 31, 170, 172

Suseela, V., Conant, R.T. et al. (2012). Effects of soil moisture on the temperature sensitivity of heterotrophic respiration vary seasonally in an old-field climate change experiment. *Global Change Biology*, 18, 336 – 348. 28, 104

Svenson, J., Santner, T., Dean, A., & Moon, H. (2014). Estimating sensitivity indices based on Gaussian process metamodels with compactly supported correlation functions. *Journal of Statistical Planning and Inference*, 144, 160 – 172. 44

Szatmari-Voicu, D. (2011). Robust M-and L-Estimators of Scale Parameter. *Communications in Statistics-Theory and Methods*, 40(22), 4065 – 4077. 40

Tebaldi, C., Smith, R., Nychka, D., & Mearns, L. (2005). Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. *Journal of Climate*, 18, 1524 – 1540. 40

Thonicke, K., Venevsky, S., Sitch, S., & Cramer, W. (2001). The role of fire disturbance for global vegetation dynamics: coupling fire into a Dynamic Global Vegetation Model. *Global Ecology and Biogeography*, 10(6), 661 – 677. 13

Thonicke, K., Prentice, I. C., & Hewitt, C. (2005). Modeling glacial-interglacial changes in global fire regimes and trace gas emissions. *Global Biogeochemical Cycles*, 19(3). 14

Tian, H., Melillo, J.M., Kicklighter, D.W. et al. (1998). Effect of interannual climate variability on carbon storage in Amazonian ecosystems. *Nature*, 396(6712), 664 – 667. 24

Tian, H., Chen, G., Liu, M. et al. (2010). Model estimates of net primary productivity, evapotranspiration, and water use efficiency in the terrestrial ecosystems of the Southern United States during 18952007. *Forest Ecology and Management*, 259(7), 1311 – 1327. 24

Tresp, V. (2000). A Bayesian committee machine. *Neural Computation*, 12(11), 2719 – 2741. 43

Tresp, V. (2001). Committee machines. *Handbook for neural network signal processing*, 1 – 18. 43

Tuomi, M., Vanhala, P., Karhu, K., Fritze, H., & Liski, J. (2008). Heterotrophic soil respiration-comparison of different models describing its temperature dependence. *Ecological Modelling*, 211(1), 182 – 190. 28, 104

Turnbull, M. H., Whitehead, D., Tissue, D. T., Schuster, W. S. F., Brown, K. J., & Griffin, K. L. (2001). Responses of leaf respiration to temperature and leaf characteristics in three deciduous tree species vary with site water availability. *Tree Physiology*, 21(9), 571 – 578. 29

van Vuuren, D.P., Edmonds J.A., Kainuma, M., Riahi, K. & Weyant J. (2011). A special issue on the RCPs. *Climatic Change*, 109, 1 – 4. 19, 20

van Wart, J., van Bussel, L.G.J, Wolf, J. & Licker, R. et al. (2013). Use of agro-climatic zones to upscale simulated crop yield potential. *Field Crops Research*, 143, 44 – 55. 27, 33

Vermeulen, S. (2014). Climate change, food security and small-scale producers: *Summary of Findings of the Fifth Assessment Report (AR5) of the Intergovernmental Panel on Climate Change (IPCC)*. 22, 31, 170

Volinsky, C. T., & Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics*, 56(1), 256 – 262. 57

Wahba, G. (1990). Spline Models for Observational Data. *Society for Industrial and Applied Mathematics*, Philadelphia. 155

Wahba, G., Johnson, D. R., Gao, F., & Gong, J. (1995). Adaptive tuning of numerical weather prediction models: Randomized GCV in three-and four-dimensional data assimilation. *Monthly Weather Review*, 123, 3358 – 3369. 155

Wallach D. (2011). Crop model calibration: A statistical perspective. *Agronomy Journal*, 103(4), 1144 – 1151. 32, 33

Wang, X., Liu, L., Piao, S., Janssens, I. A., Tang, J., Liu, W., & Xu, S. (2014). Soil respiration under climate warming: differential response of heterotrophic and autotrophic respiration. *Global Change Biology*, 20, 3229 – 3237. 29, 105

Warren R., et al. (2008). Development and illustrative outputs of the Community Integrated Assessment System (CIAS), a multi-institutional modular integrated assessment approach for modelling climate change. *Environmental Modelling and Software*, 23, 592 – 610. 3

Warren, R., Lowe, J. A., Arnell, N. W., Hope, C., Berry, P., Brown, S., ... & Vanderwal, J. (2013). The AVOID programmes new simulations of the global benefits of stringent climate change mitigation. *Climatic Change*, 120(1 – 2), 55 – 70. 2, 4

Watson, R. T., Noble, I. R., Bolin, B., Ravindranath, N. H., Verardo, D. J., & Dokken, D. J. (2000). Land use, land-use change and forestry: *A special report of the Intergovernmental Panel on Climate Change.* Cambridge University Press. 11

Webster, M.D. & Sokolov, A.P. (2000). A methodology for quantifying uncertainty in climate projections. *Climatic Change*, 46, 417 – 446. 36

Webster M., Forest, C., Reilly, J. & Babiker, M. et al. (2003). Uncertainty analysis of climate change and policy response. *Climatic Change*, 61, 295 – 320. 36

Weltzin, J.F., Loik, M.E., Schwinning, S., Williams, D.G., Fay, P., Haddad, B. et al. (2003). Assessing the response of terrestrial ecosystems to potential changes in precipitation. *BioScience*, 53, 941 – 952. 26

Whitlock, C., Shafer, S.L., & Marlon, J., (2003). The role of climate and vegetation change in shaping past and future re regimes in the northwestern US and the implications for ecosystem management. *Ecological Management*, 178, 5 – 21. 13

Williams, C. K. I. & Seeger, M. (2001). Using the Nystrom method to speed up kernel machines. In Leen, T. K., Diettrich, T. G., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, 13, 682 – 688. 43

Wilkinson, R.D. (2010a). Bayesian calibration of expensive multivariate computer experiments. Large-Scale Inverse Problems and Quantification of Uncertainty, *Ser. Comput. Stat.*, 195 – 216. 44

Wilkinson, R.D., in: Biegler et al. (Eds.) (2010b). *Large-Scale Inverse Problems and Quantification of Uncertainty.* John Wiley & Sons, New York. 41

Wold, S., Esbensen, K., Geladi, P. (1987). Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, 2, 37 – 52. Elsevier Science Publishers B.V., Amsterdam. 68

Wu, W., Verburg, P. H., & Tang, H. (2014). Climate change and the food production system: impacts and adaptation in China. *Regional Environmental Change*, 14(1), 1-5. 170

Yeniay, O. & Goktas, A. (2002). A comparison of partial least squares regression with other prediction methods. *Journal of Mathematics and Statistics*, 31, 99 – 111. 55

Young, P.C. (1999). Data-based mechanistic modelling, generalized sensitivity and dominant mode analysis. *Computer Physics Communication*, 117, 113 – 129. 35, 41

Young, P.C. and Ratto, M. (2011). Statistical emulation of large linear dynamic models. *Technometrics*, 53(1), 29 – 43. 35, 41

Yue, C., Ciais, P., Cadule, P., Thonicke, K., Archibald, S., Poulter, B., & Viovy, N. (2014). Modelling fires in the terrestrial carbon balance by incorporating SPITFIRE into the global vegetation model ORCHIDEEPart 1: Simulating historical global burned area and fire regime. *Geoscientific Model Development Discussions*, 7(2), 2377 – 2427. 14

Zhang, C. H., & Li, X. (1996). Linear regression with doubly censored data. *The Annals of Statistics*, 24(6), 2720 – 2743. 38

Zhao, L., & Hu, X. J. (2013). Estimation with right-censored observations under a semi-Markov model. *Canadian Journal of Statistics*, 41(2), 237 – 256. 40

Zhou, Y., Li, N., Dong, G., & Wu, W. (2013). Impact assessment of recent climate change on rice yields in the Heilongjiang Reclamation Area of north-east China. *Journal of the Science of Food and Agriculture*, 93(11), 2698 – 2706. 170